A SIMULATION COMPARISON OF PARAMETRIC AND NONPARAMETRIC ESTIMATORS OF QUANTILES FROM RIGHT CENSORED DATA

by

SHYAMALEE KUMARY SERASINGHE

B.Sc., University of Colombo, 1992
M.Phil., University of Peradeniya, 2005

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2010

Approved by:

Major Professor
Paul Nelson

Copyright

Abstract

Quantiles are useful in describing distributions of component lifetimes. Data, consisting of the lifetimes of sample units, used to estimate quantiles are often censored. Right censoring, the setting investigated here, occurs, for example, when some test units may still be functioning when the experiment is terminated. This study investigated and compared the performance of parametric and nonparametric estimators of quantiles from right censored data generated from Weibull and Lognormal distributions, models which are commonly used in analyzing lifetime data. Parametric quantile estimators based on these assumed models were compared via simulation to each other and to quantile estimators obtained from the nonparametric Kaplan-Meier Estimator of the survival function. Various combinations of quantiles, censoring proportion, sample size, and distributions were considered.

Our simulation show that the larger the sample size and the lower the censoring rate the better the performance of the estimates of the 5^{th} percentile of Weibull data. The lognormal data are very sensitive to the censoring rate and we observed that for higher censoring rates the incorrect parametric estimates perform the best.

If you do not know the underlying distribution of the data, it is risky to use parametric estimates of quantiles close to one. A limitation in using the nonparametric estimator of large quantiles is their instability when the censoring rate is high and the largest observations are censored.

Key Words: Quantiles, Right Censoring, Kaplan-Meier estimator

Table of Contents

List of Figures

List of Tables

Acknowledgements

I wish to express my sincere appreciation to my major Professor Dr. Paul Nelson for many reasons. First, for being my Survival Data Analysis teacher and then for undertaking to act as my major professor despite his many other academic and professional commitments. His wisdom, knowledge and commitment to the highest standards inspired and motivated me. His door was always open and he was there to meet and discuss the progress of my project. I also appreciate his continuous guidance and valuable advices throughout this project and many other difficult times I underwent during my stay at Kansas State University. I was very much inspired and enthused in working on this project by his caring supervision.

Besides my advisor, I would also like to thank Dr. James Neill, who served as a supervisory committee member. Dr. Neill was my advisor from the day I joined Kansas State University. He guided me throughout this M.Sc. program as my advisor and as the head of the department too. He was always there to listen and to give advice. He also taught me Linear models.

A special thank goes to Dr. Gary Gadbury. I initially met Dr. Gadbury as my SAS teacher, followed by the Theory of Statistics teacher and finally serving as a supervisory committee member. Thank You Dr. Gadbury, for the encouragement and guidance.

I owe my deepest gratitude to two other teachers in the past: Dr. Dallas Johnson and Dr. Juan Du.

Let me also say a big ‘Thank You’ to Dr. John Boyer the former head of the department and Pamela Schierer, the Administrative Specialist for all their help during my stay at Kansas State University,

Last, but not least, I thank my husband, Roshan for sharing his Statistics knowledge and comforting me with his infinite love.

Dedication

To my loving parents, for giving me life and educating me.

CHAPTER 1 - Introduction

Statement of Purpose

The goal of this project is to study and compare estimators of quantiles from right censored data using parametric and nonparametric methods. The performance and robustness of estimators will be investigated via simulation in terms of relative bias and relative root mean or median square error via simulation. A real data set will be used to illustrate how the estimators are obtained.

Censored Data

Censoring occurs when the value of an observation is only partially known. Censored data are frequently encountered in such areas as clinical trials, the measurement of very low levels of contaminates and in studies of the lifetimes of components. There are three basic types of censoring schemes, right, interval and left, as illustrated by the following examples. Right censoring occurs in studying the lifetimes of components where some units may still be functioning when the experiment is terminated. Interval censoring happens in a clinical trial where a subject only reports that the event of interest, such as disease relapse, occurred sometime during a relatively long period of time. Left censoring occurs when the amount of a pollutant known to be present but cannot be measured below a small threshold value. My study will only consider random right censoring as defined below. Throughout, the term *lifetime* refers to a generic positive value.

Model for Randomly Right Censored Data

Let $\underline{X} = (X_1, X_2, ..., X_n)$ be independent, identically distributed random variables representing the lifetimes of interest and having distribution F. Let $\underline{C} = (C_1, C_2, ..., C_n)$ be another family of independent, identically distributed, positive random variables independent of $\underline{X}$ and having an arbitrary continuous distribution G, called the *censoring* distribution.

Although inference about F is the goal, instead of being able to observe $\underline{X}$, we only get to observe the censored data $\{ (T_i, \delta_i) ; i = 1,..,n \}$ where $T_i = \min \{ X_i, C_i \}$; and $\delta_i = I_{(0,Ci)}(X_i)$. The probability that an observation is censored P(T=C), is given by $P[C < X] = P [C - X < 0] = P(\delta = 0)$.

Quantiles

Definition: For $0 \leq p \leq 1$, the pth *quantile* ξ_p of a distribution F is defined by

$$\xi_p = \inf \{ x, F(x) \geq p \} \equiv F^{-1}(p) \qquad (1)$$

Estimation: Let $\hat{F}$ be a right continuous estimate of F. The quantile ξ_p may then be estimated by

$$\hat{\xi}_p = \inf \{ x; \hat{F}(x) \geq p \} \equiv \hat{F}^{-1}(p). \qquad (2)$$

A distribution is uniquely determined by its quantiles. Quantiles such as the quartiles provide convenient, easy to understand summaries of a distribution. In reliability studies quantiles for p close to zero and one are also of interest.

Parametric and Non Parametric Estimators

Parametric Estimators

Here, the distribution of the lifetimes, $F = F(\cdot \mid \underline{\theta})$, is known up to a finite dimensional parameter $\underline{\theta}$. Letting $\hat{\underline{\theta}}$ denote an estimator of $\underline{\theta}$ obtained from right censored data, the quantile estimator in (2) becomes

$$\xi_p(F) = \inf \{ x, F(x \mid \hat{\underline{\theta}}) \geq p\} \equiv F^{-1}(p \mid \hat{\underline{\theta}}). \qquad (3)$$

My study will use two widely used choices for F, Weibull and Lognormal distributions.

Non Parametric Estimator

Here, $\hat{\xi}_p(F) = \inf \{ x, \hat{F}(x) \geq p\}$, where $\hat{F} \equiv 1 - \hat{S}$ and $\hat{S}$ is the Kaplan-Meier Estimator of the survivor function S. The survivor function (reliability function) is the probability that the system will survive beyond a specific time

$$\begin{aligned} S(t) &= P[\, T > t \,] \\ &= 1 - P[\, T \leq t \,] \\ &= 1 - F(t) ; \qquad t > 0 \end{aligned}$$

The Kaplan-Meier estimator $\hat{S}(t) = 1 - \hat{F}(t)$ can be viewed as the nonparametric maximum likelihood estimator of the survivor function based on censored data. No parametric form of F is required. We then have :

$$\hat{\xi}_p = \inf \{ x; \hat{F}(x) \geq p \}$$

$$= \inf \{ x, \hat{S}(x) \leq (1 - p) \}$$

CHAPTER 2 - Motivation

Common Practice

Engineering researchers tend to favor the use of parametric methods for analyzing lifetime data with censoring. In contrast, the Medical researchers tend to favor non- parametric methods since they are reluctant to base their analyses on assumed types of distributions.

Failure Time Distributions

There are several ways of specifying an absolutely continuous probability distribution supported on the positive reals: The Probability Density Function $f(\cdot)$ (PDF), The Cumulative Distribution function $F(\cdot)$ (CDF), The Survival Function $S = 1 - F$, and The Hazard Function(Rate) $h(\cdot)$. The Survival Function and the Hazard Rate are heavily used in survival analysis. There are many types of survivor curves, all with some basic properties: monotone, nonincreasing functions equal to one at zero and zero at infinity. The hazard function can be increasing, decreasing, constant, bathtub-shaped, or hump-shaped.

The Hazard Function

The hazard function (also known as the conditional failure rate in reliability, the force of mortality in demography, or the age specific failure rate in epidemiology) for a continuous random variable is the ratio of the probability density function to the survival function and for S(x) > 0 is given by

$$h(x) = \lim_{\Delta x \to 0} \frac{P(x < X \le x + \Delta x \mid X > \Delta x)}{\Delta x}$$

$$= \frac{f(x)}{S(x)} \quad .$$

The only restrictions on h(x) are that it be nonnegative and $e^{-\int_0^x h(t)dt} \to 0 \quad as \quad x \to \infty$. A survivor function gives the probability of survival as a function of time. The hazard function

gives the instantaneous probability of failure given survival up to a given time. For example, for an exponential distribution with scale parameter $\lambda > 0$, for $x > 0$,

$$f(x) = \frac{1}{\lambda} e^{-\left(\frac{x}{\lambda}\right)}, \; S(x) = e^{-\left(\frac{x}{\lambda}\right)}, \text{ and } \; h(x) = \frac{1}{\lambda}, \text{ is a constant.}$$

Raising an exponential random variable to a positive power leads to the Weibull distribution, which is flexible enough to accommodate three hazard rates, decreasing (when shape parameter <1), constant (when shape parameter =1, and identical to the exponential distribution) and increasing (when shape parameter >1). See Figure 2.1. The Weibull distribution is widely used in reliability and life data analysis.

The probability density function of a Weibull distribution with scale parameter $\lambda > 0$, shape parameter $\beta > 0$, x > 0, is given by

$$f_x(x;\beta,\lambda) = \begin{cases} \left(\frac{\beta}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\beta-1} e^{-\left(\frac{x}{\lambda}\right)^{\beta}} & ; \; x \geq 0 \\ 0 & ; \; x < 0 \end{cases},$$

$$S(x) = e^{-\left(\frac{x}{\lambda}\right)^{\beta}},$$

$$h(x) = \left(\frac{\beta}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{(\beta-1)}$$

The Lognormal distribution is popular because of its relationship to the Normal distribution. Specifically if X is lognormal, log(X) is normal. Further, the lognormal hazard function has non-monotone behavior. It increases initially, then decreases and eventually approaches zero. See Figure 2.2. This means that lifetimes with a lognormal distribution can have a higher rate of failing as they age for some period of time, but after survival to a specific age, the rate of failure decreases as time increases.

For a Lognormal distribution with parameters $\mu, \sigma > 0, x > 0$,

$$f_x(x;\mu,\sigma) = \begin{cases} \left(\frac{1}{x\sigma\sqrt{2\pi}}\right) e^{-\left(\frac{(\ln x-\mu)^2}{2\sigma^2}\right)} & ; \ x > 0 \\ 0 & ; \ x \le 0 \end{cases}$$

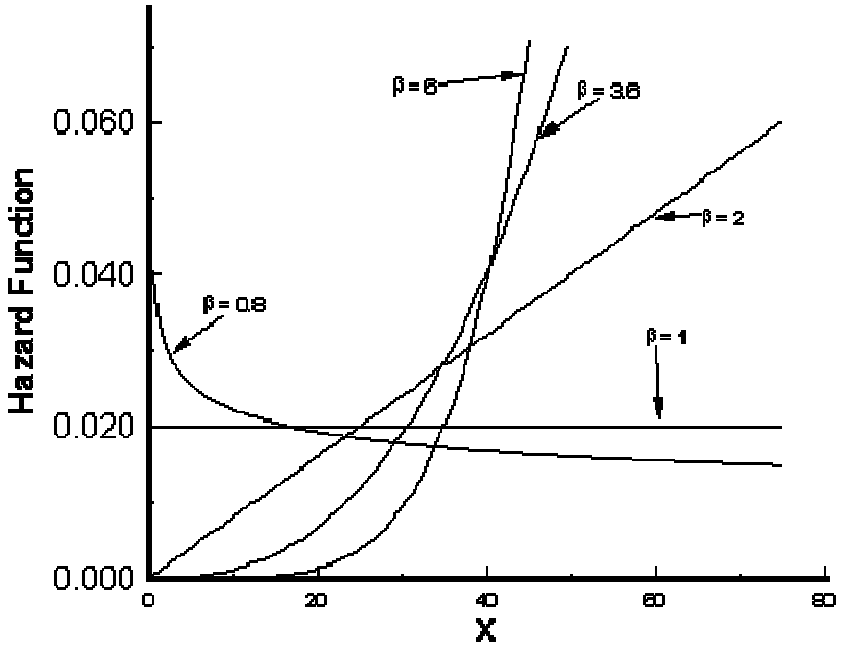
$$S(x) \quad = \quad 1 \quad - \quad \Phi\left[\frac{\ln(x)-\mu}{\sigma}\right].$$

where Φ is the distribution function of a standard normal distribution.

Hence,

$$h(x) \quad = \quad \frac{\left(\frac{1}{x\sigma\sqrt{2\pi}}\right) e^{-\left(\frac{(\ln x-\mu)^2}{2\sigma^2}\right)}}{1 \quad - \quad \Phi\left[\frac{\ln(x)-\mu}{\sigma}\right]},$$

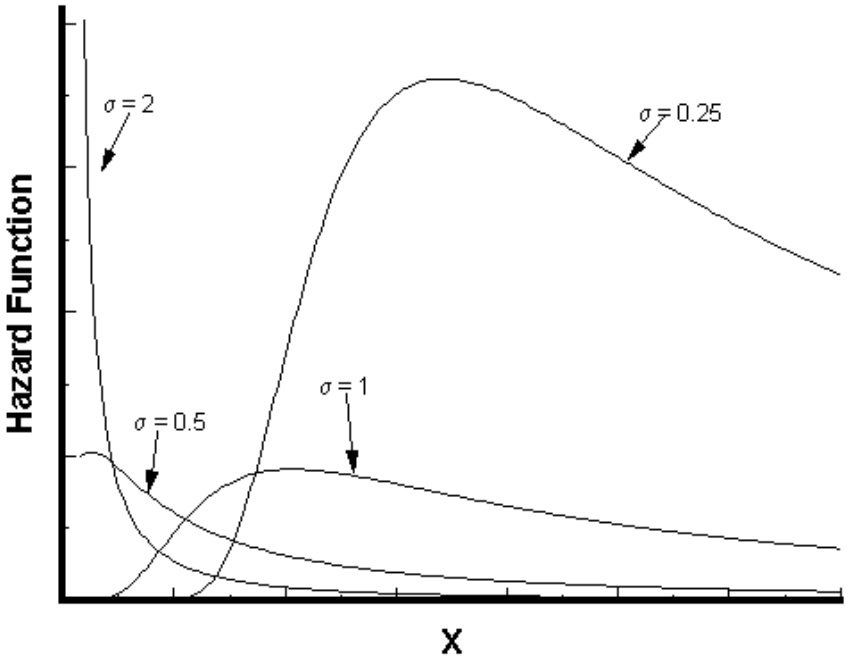
Figure 2.1 Weibull Hazzard Functions when λ = 1



From site:

http://www.engineeredsoftware.com/nasa/weibull.htm

Figure 2.2 Lognormal Hazard Functions when μ = 0



From site:

http://www.engineeredsoftware.com/nasa/Lognormal.htm

CHAPTER 3 - Theoretical Background

Parametric Likelihood Construction for Censored Data

Let the density $f(x \mid \underline{\theta})$ and survivor function $S(x \mid \underline{\theta})$ depend on an unknown parameter $\underline{\theta}$. Throughout we assume that the distribution of the censoring $\{C_i\}$ times do not depend on $\underline{\theta}$. According to Klein & Moeschberger (2003), "An observation corresponding to an exact event time provides information on the probability that the event's occurring at this time, which is approximately equal to the density function of X at this time. For a right-censored observation all we know is that the event time is larger than this time, so the information is the survival function evaluated at the on study time"

Accordingly, we obtain a likelihood $L(\underline{\theta})$ given by

$$L(\theta) = \prod_{i \in D} f(x_i \mid \underline{\theta}) \prod_{i \in R} S(C_r \mid \underline{\theta}),$$

where D is the set of uncensored times and R is the set of right censored observations. A critical assumption made here is that the lifetimes and censoring times are independent. For a right censored sample of (T_i, δ_i), i = 1, … , n, the likelihood function can also be expressed as

$$L(\underline{\theta}) = \prod_{i=1}^{n} \left[f(t_i \mid \underline{\theta})\right]^{\delta_i} \quad \left[S(t_i \mid \underline{\theta})\right]^{1-\delta_i}. \qquad (3.1)$$

Accelerated Failure-Time Model

The Accelerated Failure Time (AFT) model relates covariates linearly to the logarithm of the survival time. Let X denote the time to the event and $\underline{Z}$ a vector of fixed time, explanatory covariates. The AFT model states that the survival function of an individual with covariate $\underline{Z}$ at time x is the same as the survival function of an individual with fully specified baseline survival

function S_0 and hazard function h_0 at a time $x\,e^{(\underline{\theta}^t \underline{Z})}$, where $\underline{\theta}^t = (\theta_1,...,\theta_p)$ is a vector of regression coefficients. The AFT models is defined by the relationship

$$S(x \mid \underline{\theta}, \underline{Z}) = S_0[e^{\underline{\theta}^t \underline{Z}} x],\ h(x \mid Z, \theta) = \exp(\underline{\theta}^t Z) h_0(\exp(\underline{\theta}^t Z x)),\ \textit{for all}\quad x.$$

The factor $e^{\underline{\theta}^t \underline{Z}}$ is called an acceleration factor telling the investigator how a change in covariate values changes the time scale from the baseline time scale. The likelihood for data $\{(t_i, \delta_i, \underline{z}_i)\}$ can then be written as

$$L(\underline{\theta}) = \prod_{i=1}^{n} \left[f(t_i \mid \underline{Z}_i, \underline{\theta}\right]^{\delta_i} \quad \left[S(t_i \mid \underline{Z}_i, \underline{\theta})\right]^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \left[h(t_i \mid \underline{Z}_i, \underline{\theta}\right]^{\delta_i} \quad \left[S(t_i \mid \underline{Z}_i, \underline{\theta})\right]^{1-\delta_i},$$

where $\delta_i = I_{(0,C_i)}(X_i)$. Covariates are not included in my study.

Kaplan - Meier Estimator

To obtain the non-parametric estimator of quantiles, we use the Kaplan-Meier estimator of the survivor function, is also known as the Product Limit Estimator. For the purpose of illustration, consider a set of right censored data. Such data are represented using two numbers,

T - Time under observation

δ - Indicator of failure/censoring

A key assumption we make here is that the potential censoring time is unrelated to the potential event time. Possible ties in the data set are also allowed. The events occur at D, distinct times $t_1<t_2<\ldots<t_D$. At time t_i there are d_i events observed. Let Y_i be the number of individuals who are at risk at time t_i. Specifically, Y_i is the number of individuals who are alive at t_i or experience the event of interest at t_i. Now, d_i/Y_i provides an estimate of the conditional probability that an individual who survives adjust prior to time t_i experiences the event at time t_i. This is the basic quantity from which we will construct estimators of the survival function and then the quantiles. For all values of t in the range where there is data

$$\hat{S}(t) = \begin{cases} 1 & ;t < t_1 \\ \Pi_{t_i \le t}\left[1 - \frac{d_i}{Y_i}\right] & ;t_1 \le t \end{cases}$$

, $t_1 < t_2 < \ldots < t_D$

The Kaplan-Meier estimator is a right continuous step function with downward jumps at the observed times. It is not defined for values beyond the largest observed failure time.

SAS Procedures Proc Lifereg and Proc Lifetest

SAS Procedure used for Parametric Estimators

The LIFEREG procedure in SAS fits parametric models to failure time data that can be right, left, or interval censored. It allows covariates to be associated with each unit. LIFEREG can also be used in studies like mine which do not incorporate covariates. The LIFEREG procedure, described below, estimates the parameters by maximum likelihood using a Newton-Raphson algorithm.

PROC LIFEREG < options > **;**
MODEL response=independents < **/** options > **;**
BY variables **;**
CLASS variables **;**
OUTPUT < **OUT=**SAS-data-set >
keyword**=**name < : : : keyword**=**name >
< options > **;**
WEIGHT variable **;**

MODEL Statement

label:> **MODEL** response<*censor(list)>=independents </ options > **;**

censor is a binary variable indicating whether the observation is censored or not. Valid values of the variable censor are 0 (yes) and 1 (no). The term ‘ independents’ refer to covariates.

The DISTRIBUTION = option can be used to specify distribution type for failure time in the MODEL statement. If the MODEL statement does not specify the DISTRIBUTION= option, the LIFEREG procedure fits the default type 1 extreme value distribution using log(.) as the response. This is equivalent to fitting the Weibull distribution.

We will use the designation DISTRIBUTION = lognormal for fitting a lognormal distribution to the data.

SAS Procedure used for Nonparametric Estimators

The LIFETEST procedure computes nonparametric estimates of the survival function and hypothesis tests for the equality of 2 or more distributions. You can request either the product-limit (Kaplan-Meier) or the life-table (actuarial) estimate of the distribution. PROC LIFETEST computes nonparametric tests to compare the survival curves of two or more groups.

The following statements are available in PROC LIFETEST:

PROC LIFETEST < options > **;**

TIME variable < *censor(list) > **;**

BY variables **;**

FREQ variable **;**

ID variables **;**

STRATA variable < (list) > < ... variable < (list) > > **;**

SURVIVAL options **;**

TEST variables **;**

Some of these options include:

METHOD=type

specifies the method used to compute the survival function estimates. Valid values for type are as follows.

PL | KM

specifies that product-limit (PL) or Kaplan-Meier (KM) estimates are computed.

ACT | LIFE | LT

specifies that life-table (or actuarial) estimates are computed.

By default, METHOD=PL.

Data Generation

Probability Integral Transformation

Recall that the Weibull distribution is a continuous distribution with probability density function

$$f_x(x;\beta,\lambda)=\begin{cases}\left(\frac{\beta}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\beta-1}e^{-\left(\frac{x}{\lambda}\right)^{\beta}} & ;\ x\geq 0\\ 0 & ;\ x<0\end{cases}$$

for shape parameter $\beta>0$, scale parameter $\lambda>0$. The cumulative distribution function is

$$F_x(x;\beta,\lambda)=\begin{cases}1-e^{-\left(\frac{x}{\lambda}\right)^{\beta}} & ;\quad x\geq 0\\ 0 & ;\ x<0\end{cases}$$

For data generation from a Weibull distribution we applied the inverse of the probability integral transformation to convert random variables from a uniform distribution to random variables from a Weibull distribution.

Specifically let U ~ U(0,1), Since $F_x(x;\beta,\lambda)=1-e^{-\left(\frac{X}{\lambda}\right)^{\beta}} \sim U(0,1)$, solving for 'X', we obtain

$$e^{-\left(\frac{X}{\lambda}\right)^{\beta}}=1-U$$

$$-\left(\frac{X}{\lambda}\right)^{\beta}=\log(1-U)$$

$$\left(\frac{X}{\lambda}\right)^{\beta}=-\log(1-U)$$

$$\left(\frac{X}{\lambda}\right)=\left(-\log(1-U)\right)^{\frac{1}{\beta}}.$$

Hence,

$$X = \lambda\left(-\log(1-U)\right)^{\frac{1}{\beta}},$$

or equivalently,

$$X = \lambda\left(-\log(U)\right)^{\frac{1}{\beta}}$$

has a Weibull distribution.

Box-Muller Algorithm

The Box-Muller algorithm was used to generate Lognormal data. The Box–Muller algorithm given in Box and Muller (1958) [http://en.wikipedia.org/wiki/Box%E2%80%93 Muller_transform–cite note–0] is a method for generating pairs of independent standard normally distributed (zero expectation, unit variance) random numbers, given a source of independent, uniformly distributed random numbers. Suppose U_1 and U_2 are independent random variables that are uniformly distributed in the interval (0,1]. Let

$$Z_1 = \sqrt{-2\ln U_1}\cos(2\pi U_2), \text{ and } Z_2 = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

where Z_1 and Z_2 are independent standard normal random variables. Then we use $X = \exp(\mu + \sigma Z)$, Z = Z_1 Z_2 , to generate lognormal variates. The p^{th} quantile ξ_p, for the two parametric distributions were computed using the following:

$$\text{Weibull } \xi_p = \lambda\left[\left(-\log(1-p)\right)\right]^{\frac{1}{\beta}},$$

$$\text{Lognormal } \xi_p = \exp(\mu + \Phi^{-1}(p)\sigma).$$

Generating Censored Observations

Generating censored data requires specifying both the lifetime and censoring distributions, denoted F and G respectively. As stated above, based on their wide spread use, we generated Weibull and lognormal lifetimes dependent on an unknown parameter θ. One of our main goals is to study the impact of censoring rate $\pi = \psi(F, G)$ on the estimation of lifetime quantiles. Our simulation study specifies F, θ and π, making G dependent on θ, which would violate one of our assumptions. To avoid this issue and simplify our simulation study, we used the following modified algorithm for generating right censored data. Independent of the lifetimes $\{ X_i \}$, a sequence of independent Bernoulli random variables $\{ \delta_i \sim B(1, 1-\pi) \}$ was used to denote whether a lifetime is censored or not. This series was generated using the function ranbin(), [http://support.sas.com/documentation/cdl/en/lrdict/63026/HTML/default/viewer.htm#/documentation/cdl/en/lrdict/63026/HTML/default/a000202883.htm]

Note that this algorithm generates the same likelihood as given in (3.1).

CHAPTER 4 - Simulation Study

Methodology

The following algorithm was used to carry out the simulation study.

Step 1 : Specify quantiles $\underline{\xi} = \{\xi_{p_i}, i = 1,2,...,k\}$, censoring proportions $\underline{\pi} = \{\pi_i, i = 1,2,...,m\}$, sample size n and distribution F.

Step 2 : Fix ξ_p, π, n and F=F$^{\#}$.

Step 3 : Generate $\{ X_i, i = 1,2,...,n\}$, independent, each with distribution F$^{\#}$.

Step 4 : Generate $\{ \delta_i \sim B(1, 1-\pi) \}$ independent Bernoulli random variables.

Step 5 : Specify $T_i \equiv X_i$ as being a detection only if $\delta_i = 1$, resulting in data $\{(T_i, \delta_i), i = 1,2,\ldots,n\}$

Step 6 : Consider three estimators of ξ_p, $\{ \hat{\xi}_{p,KM}, \hat{\xi}_{p,W}, \hat{\xi}_{p,LN} \}$ based on the following three estimators of F$^{\#}$:

$\hat{F}_{KM}$: Kaplan-Meier Estimator, quantile estimator $\hat{\xi}_p(KM)$

$\hat{F}_W$: $F_W(\cdot, \hat{\boldsymbol{\theta}})$, where $F_W(\cdot, \boldsymbol{\theta})$ is a Weibull distribution function with

Parameters $\boldsymbol{\theta} = \{\lambda, \beta\}$, $\hat{\boldsymbol{\theta}}$ is the MLE of $\hat{\theta}$, quantile estimator $\hat{\xi}_p(W)$

$\hat{F}_{LN}$: $F_{LN}(\cdot, \hat{\boldsymbol{\theta}})$, where $F_W(\cdot, \boldsymbol{\theta})$ is a Lognormal distribution function with

Parameters $\boldsymbol{\theta} = \{\mu, \sigma\}$, $\hat{\boldsymbol{\theta}}$ is the MLE of $\hat{\theta}$, quantile estimator $\hat{\xi}_p(LN)$

Step 7 : Independently repeat (1)-(4) N times obtaining quantile estimators $\{\hat{\xi}_{KM}, \hat{\xi}_{W}, \hat{\xi}_{LN}\}_{i=1}^{N}$

Step 8 : Estimate the mean and mean square error of these estimators.

Step 9 : Summarize the results and draw comparative conclusion about the performance of the estimators when the correct and incorrect parametric and nonparametric estimators are used.

Note that if $F^{\#}$ is a lognormal distribution, $\hat{\xi}_p(W)$ is computed using a wrong likelihood. Likewise, $\hat{\xi}_p(LN)$ is obtained from an incorrect likelihood when $F^{\#}$ is a Weibull distribution.

Setting up the Parameters

Quantiles $\underline{\xi} = \{\xi_{p_i}, i = 1,2,...,k\}$, censoring proportions $\underline{\pi} = \{\pi_i, i = 1,2,...,m\}$, sample size n and parameter setting of the distribution F were the things to be set.

Censoring Proportions

In an attempt to match what is seen in the real world we selected 0.00, 0.05, 0.25, and 0.50 as our censoring proportions.

Quantiles

Since extreme quantiles such as the 5 th and 95 th are typically of interest in survival data analysis and in the interest of time, only those two were considered.

Sample Sizes

25 , 40 and 60 were used to represent small, middle and large sample sizes.

Iterations

Stick to 1000 if possible or else will consider a smaller number.

Distributions

We simulated data from two lifetime distributions, Weibull and Lognormal and considered four different combinations of parameters for each.

Considering these parameter selections altogether we ended up with following combinations. Altogether 4(censoring proportions) x 2(quantiles) x 3(sample sizes) x 4(different shapes) x 2(distributions) =192 simulations were done.

Problems Encountered and How We Overcame Them

The simulation ran smoothly for a smaller number of iterations. But when attempting 1000 iterations the computer ran out of resources such as the capacity of the log window. So we had to clean the log window, which slowed down the execution of the program.

When estimating the quantile close to 1, i.e. the 95 th percentile, using the non-parametric method, we ended up with missing values in some cases where the largest observation was censored. In such cases we considered only the observations with all three estimators defined. The observations with missing values were discarded. The following tables give the number of missing values for each case. As the censoring proportion increases more missing values were observed. Further, we can see that as the sample size increases the number of missing values observed was decreased.

Table 4.1 Number of Missing Values out of 1000 iterations for Weibull Data

Censoring Proportion		5%			25%			50%		
		Sample Size			Sample Size			Sample Size		
Beta	Lambda	**25**	**40**	**60**	**25**	**40**	**60**	**25**	**40**	**60**
0.5	1	11	1	1	228	171	93	507	501	496
1	1	7	1	1	262	164	81	494	508	505
1.5	1	9	4	0	260	150	101	489	478	498
5	1	8	2	5	231	165	94	497	492	478

Table 4.2 Number of Missing Values out of 1000 iterations for Lognormal Data

Censoring Proportion		5%			25%			50%		
		Sample Size			Sample Size			Sample Size		
Sigma	Mu	**25**	**40**	**60**	**25**	**40**	**60**	**25**	**40**	**60**
10	0	14	9	0	228	156	94	503	486	495
1.5	0	15	2	1	228	157	95	505	486	495
1	0	14	2	0	230	156	93	503	486	495
0.25	0	14	5	0	228	157	93	506	486	495

For some datasets there was a considerable amount of extreme outliers among the 95th percentile estimates, as high as 10% with some datasets. These very large estimated values greatly skewed the estimates of the mean square errors. So we decided to look at the medians of $\{ (\hat{\xi}_j - \xi)^2 \}$ instead of the means. To keep the uniformity of the analysis we considered the medians for all the 95th percentile estimates.

Simulation Output

Recall the parameters we set and used for the simulation:

ξ_p - percentile (5th or 95th)

π - censoring proportion (5%, 25%, and 50%)

n - sample size (25, 40 and 60)

F - underline distribution (Weibull and Lognormal)

In order to consider the different shapes of the two underline distributions

β - Weibull shape parameter (0.5, 1.0, 1.5 and 5.0)

λ - Weibull scale parameter (always 1)

σ - Lognormal shape parameter (0.25, 1.0, 1.5 and 10)

μ - Lognormal location parameter (always 0)

From our 1000 iterations we obtained $\{\hat{\xi}_{KM}, \hat{\xi}_{W}, \hat{\xi}_{LN}\}_{i=1}^{N}$, where N = 1000. Then the mean, bias, relative bias, mean square error and relative root mean square error of the estimators were computed as follows.

$$\bar{\hat{\xi}}_{KM} = \frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_{KM}(i)$$

$$\text{Bias}_{\text{KM}} = \hat{\xi}_{KM}(i) - \xi_p, \quad \text{Relative Bias}_{\text{KM}} = \frac{\left|\bar{\hat{\xi}}_{KM} - \xi_p\right|}{\xi_p}$$

$$\hat{MSE}_{KM} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\xi}_{KM}(i) - \xi_p)^2, \text{ Estimted Re}\textit{lative root } MSE_{KM} = \frac{\sqrt{\hat{MSE}_{KM}}}{\xi_p}$$

$$\bar{\hat{\xi}}_{W} = \frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_{W}(i)$$

$$\text{Bias}_{\text{W}} = \hat{\xi}_{W}(i) - \xi_p, \quad \text{Relative Bias}_{\text{W}} = \frac{\left|\bar{\hat{\xi}}_{W} - \xi_p\right|}{\xi_p}$$

$$\hat{MSE}_{W} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\xi}_{W}(i) - \xi_p)^2, \text{ Estimated Re}\textit{lative root } MSE_{W} = \frac{\sqrt{\hat{MSE}_{W}}}{\xi_p}$$

$$\bar{\hat{\xi}}_{LN} = \frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_{LN}(i)$$

$$\text{Bias}_{\text{LN}} = \hat{\xi}_{LN}(i) - \xi_p, \quad \text{Relative Bias}_{\text{LN}} = \frac{\left|\bar{\hat{\xi}}_{LN} - \xi_p\right|}{\xi_p}$$

$$\hat{MSE}_{LN} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\xi}_{LN}(i) - \xi_p)^2, \text{ Estimated Re}\textit{lative root } MSE_{LN} = \frac{\sqrt{\hat{MSE}_{LN}}}{\xi_p}$$

Application to a Real Data Set

The times to death for patients with cancer of the tongue is given in (Klein & Moeschberger, p. 12) and in my Appendix A. We considered only the Aneuploid Tumors data set with 52 observations, including 21 right censored observations. The censoring proportion here is thus about 40%. The dataset, SAS code, and SAS output can be found in Appendix A. Since the largest observation is censored here, the Kaplan-Meier estimator presented in Figure 4.1 is never zero and the nonparametric estimate of the 95th percentile cannot be computed here. In such situations, a common practice is to redefine the largest observation as a lifetime. The resulting Kaplan-Meier estimator is presented in Figure 4.2. Note the large difference among the estimators in this case in Table 4.4.

Figure 4.1 Plot of Kaplan-Meier Estimator – When the Largest Observation is Censored

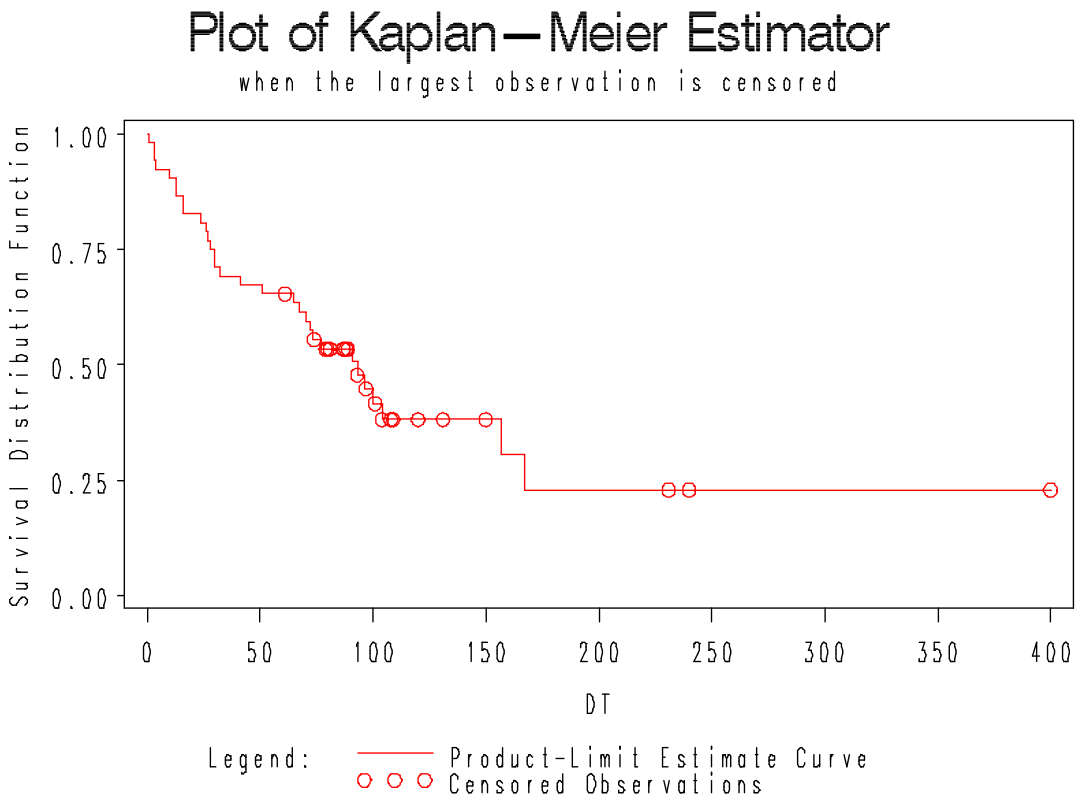


Table 4.3 Results when the largest observation is censored

	5th Percentile	95th Percentile
Kaplan-Meier	3	-
Assuming Weibull	1.475629	195.7018
Assuming Lognormal	5.167433	1457.13

Figure 4.2 Plot of Kaplan-Meier Estimator – When the Largest Observation is Uncensored

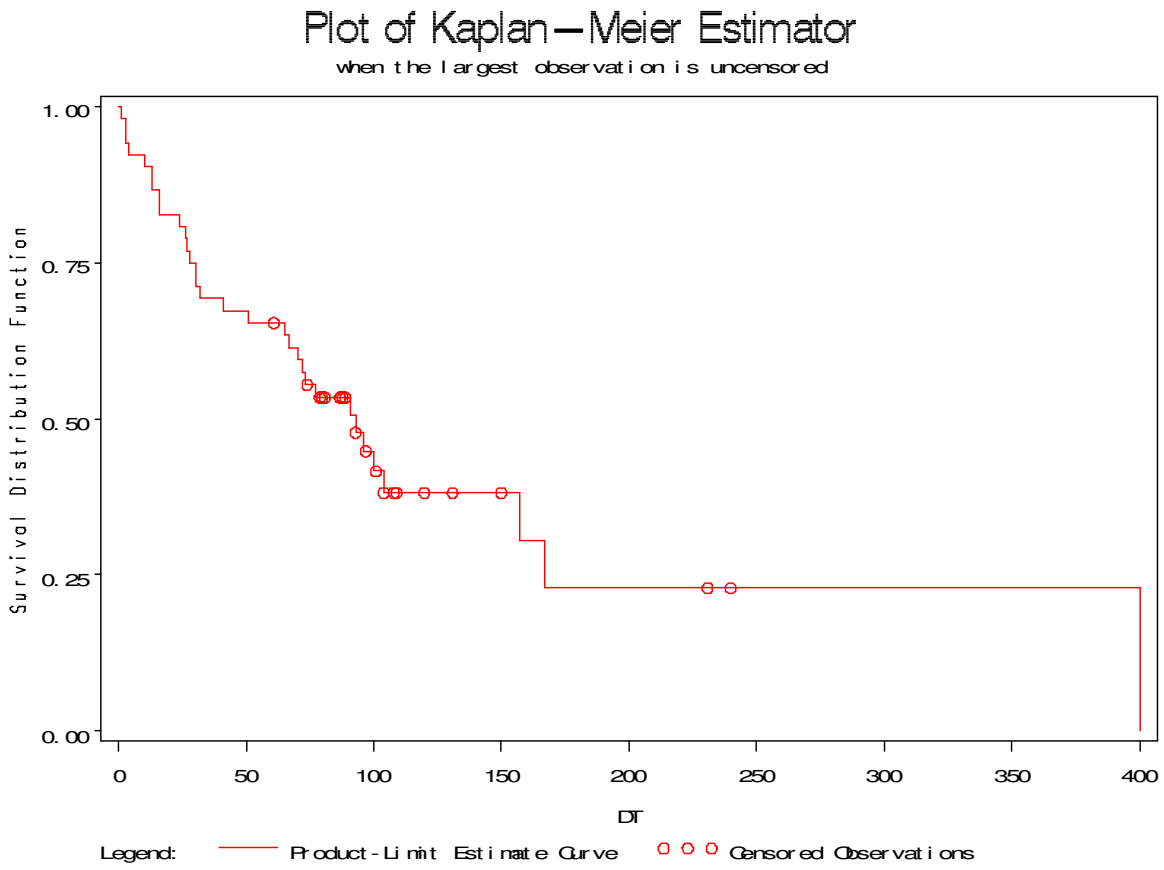


Table 4.4 Results when the largest observation is uncensored

	5^{th} Percentile	95^{th} Percentile
Kaplan-Meier	3	400
Assuming Weibull	1.747468	182.9595
Assuming Lognormal	5.413027	1286.385

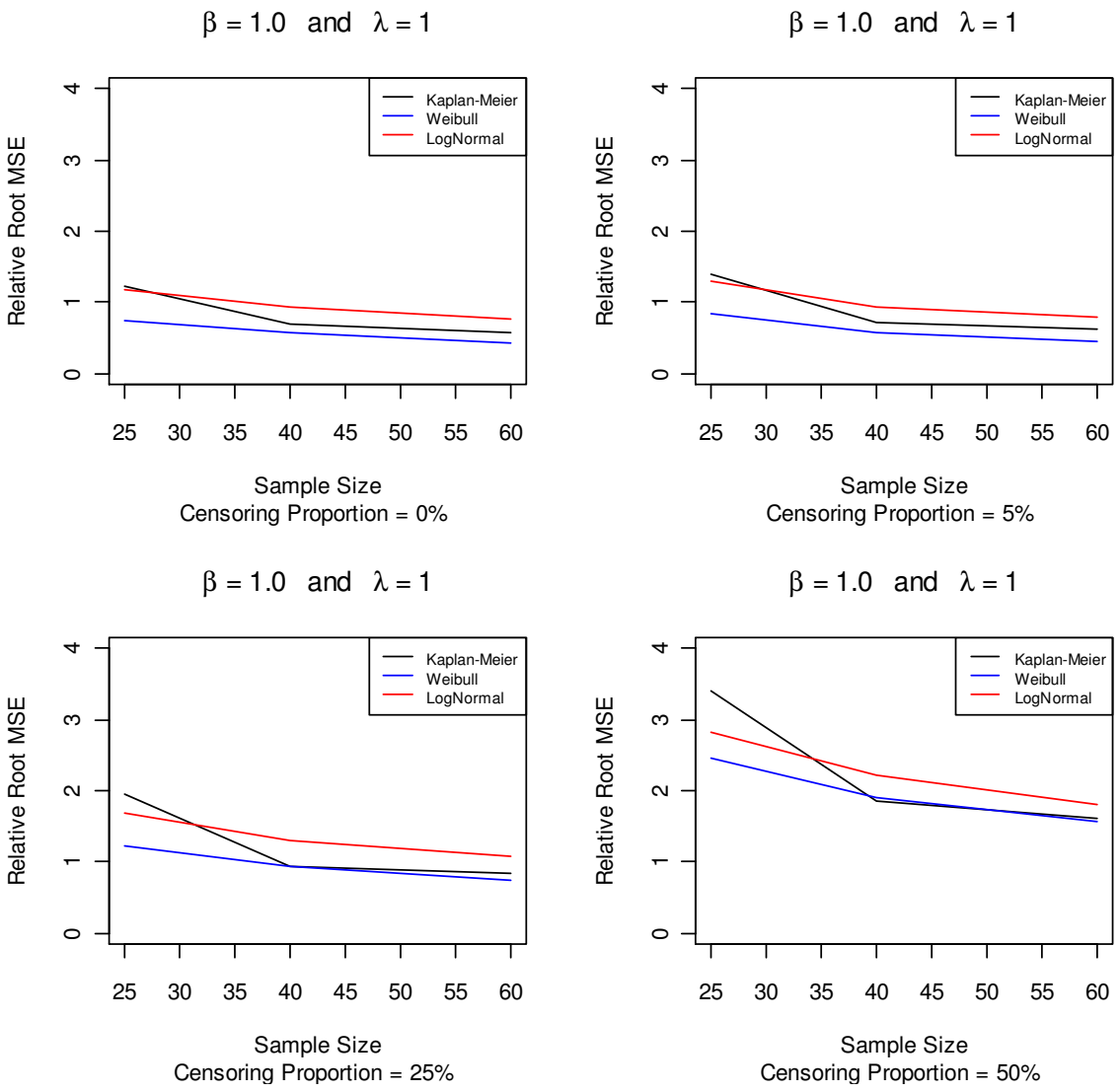
CHAPTER 5 - Results & Discussion

Relative Root Mean/Median Square Error and Relative Bias

Figures 5.1 and 5.2 present plots of estimated relative root mean square error vs sample size under various levels of censoring proportions for Weibull data for the specified shape and scale parameters. Here the blue profile represents the correct Weibull estimates and the red profile the 'incorrect' lognormal estimates. The black profile represents the nonparametric Kaplan-Meier estimates.

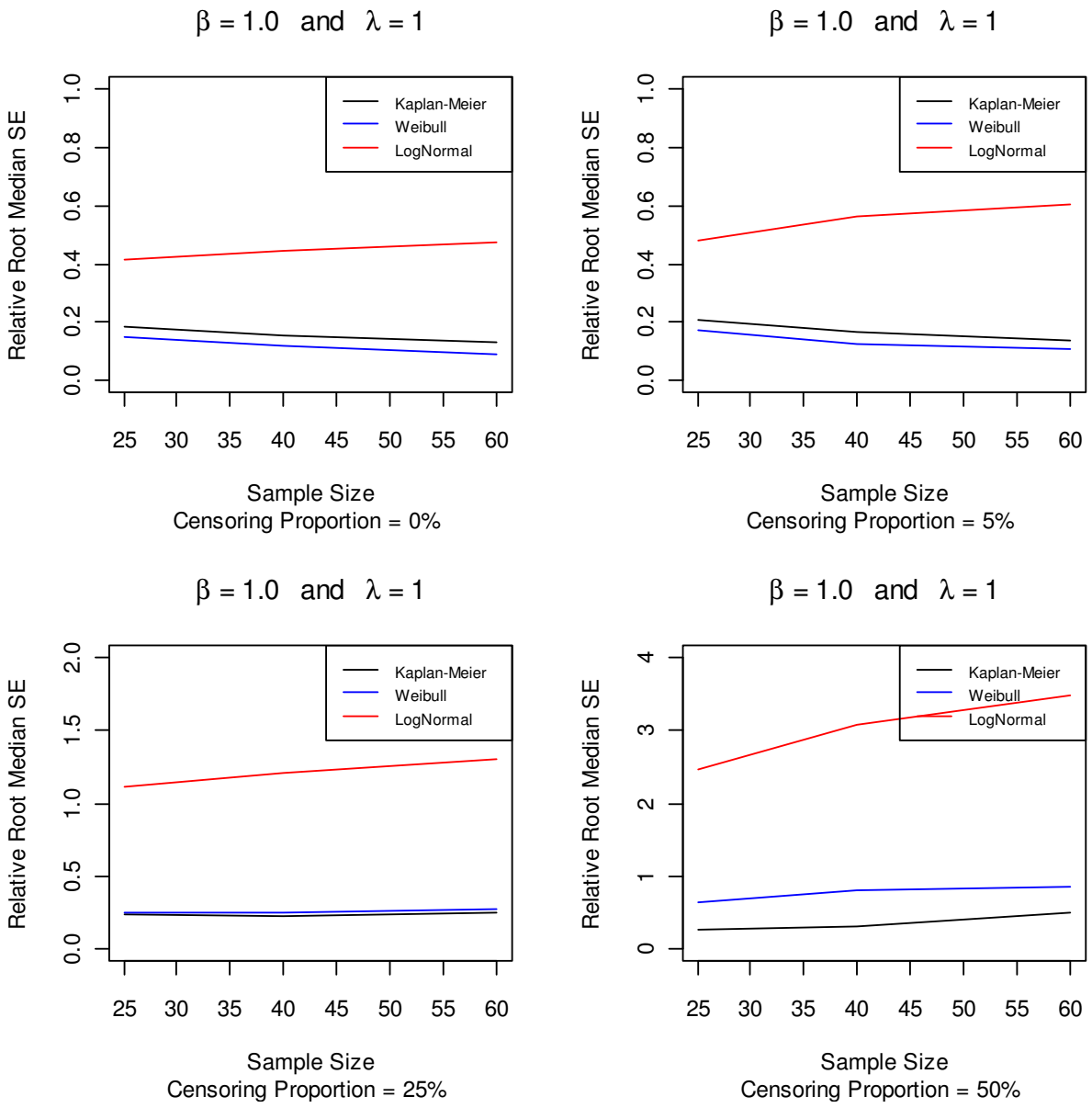
As an arbitrary but useful benchmark, I consider estimators with relative root mean/median square errors and/or relative biases more than one to be unsatisfactory. Recall that the median square error instead of the mean square error was used in assessing estimators of the 95^{th} percentile.

Figure 5.1 **5^{th} Percentile Relative Root Mean Square Error Plots for Weibull Data with $\hat{a} = 1.0$ and $\ddot{e} = 1$**



Overall, estimates of the 5^{th} percentile in Figure 5.1 improve with increasing sample size and decreasing censoring rate and all three estimates are very close to one another. However in most cases none of the above estimates can be considered as satisfactory since the relative root mean square error is closer to or above one.

Figure 5.2 **95th Percentile Relative Root Median Square Error Plots for Weibull Data with $\hat{a} = 1.0$ and $\ddot{e} = 1$**

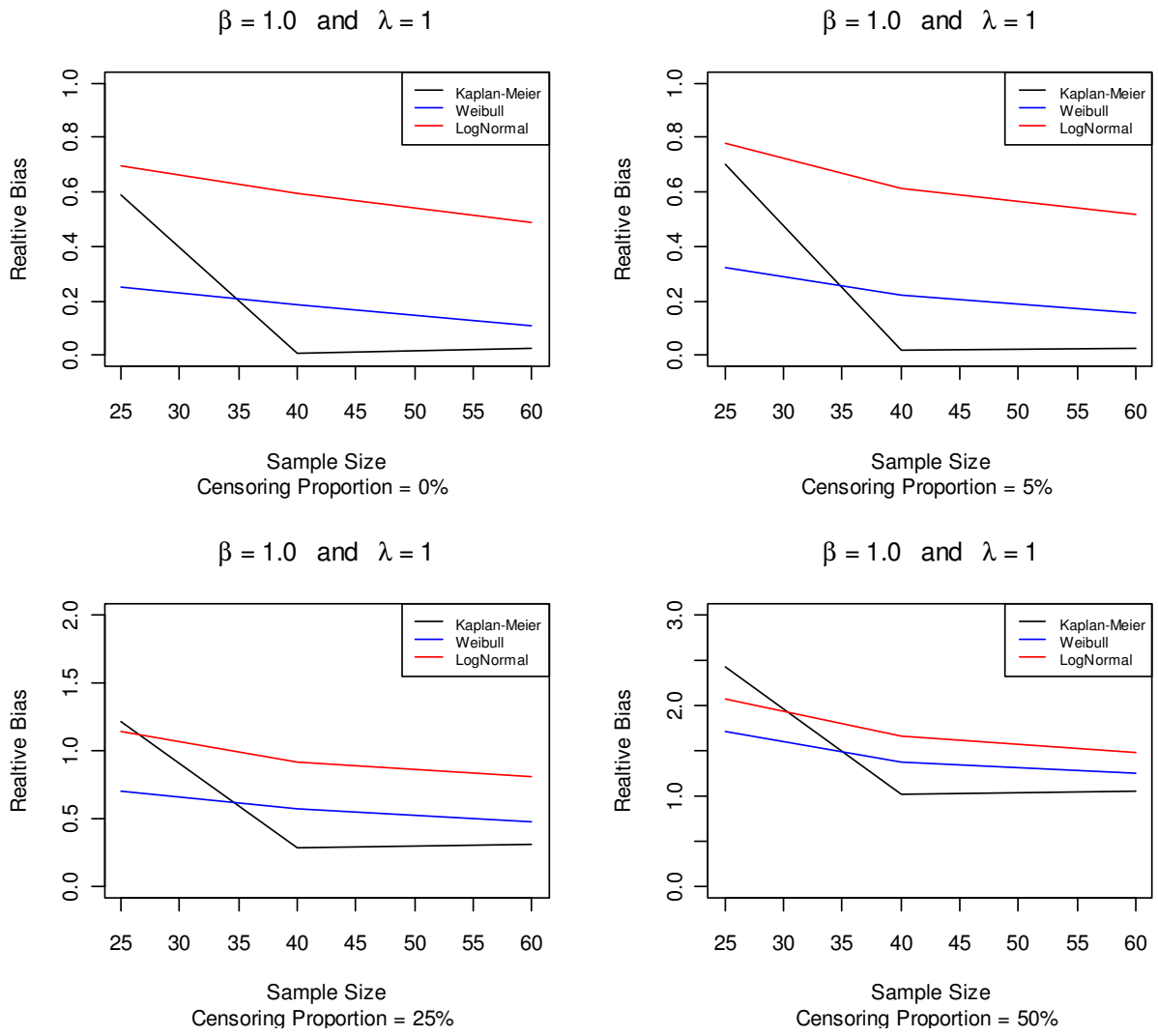


The vertical scales in the panels of Figure 5.2 are not the same in order to accommodate the vastly different relative root median square errors that resulted from estimating the 95th percentile of Weibull data. Here, estimates from the incorrect lognormal analysis are much worse than the other two, whose profiles are almost identical. Overall, relative root median square error improves with decreasing censoring rate and is relatively stable across sample sizes. When we consider the magnitude of the relative root median square error both the correct parametric estimator and the nonparametric estimator behave well, but the incorrect parametric estimator

yields median square errors far too above one and are again unsatisfactory, especially with increasing censoring rates.

Figures 5.3 and 5.4 present plots of estimated relative bias vs sample size under various levels of censoring proportions for Weibull data for the specified shape and scale parameters. Again, the blue profile represents the correct Weibull estimates and the red profile the 'incorrect' lognormal estimates. The black profile represents the nonparametric Kaplan-Meier estimates.

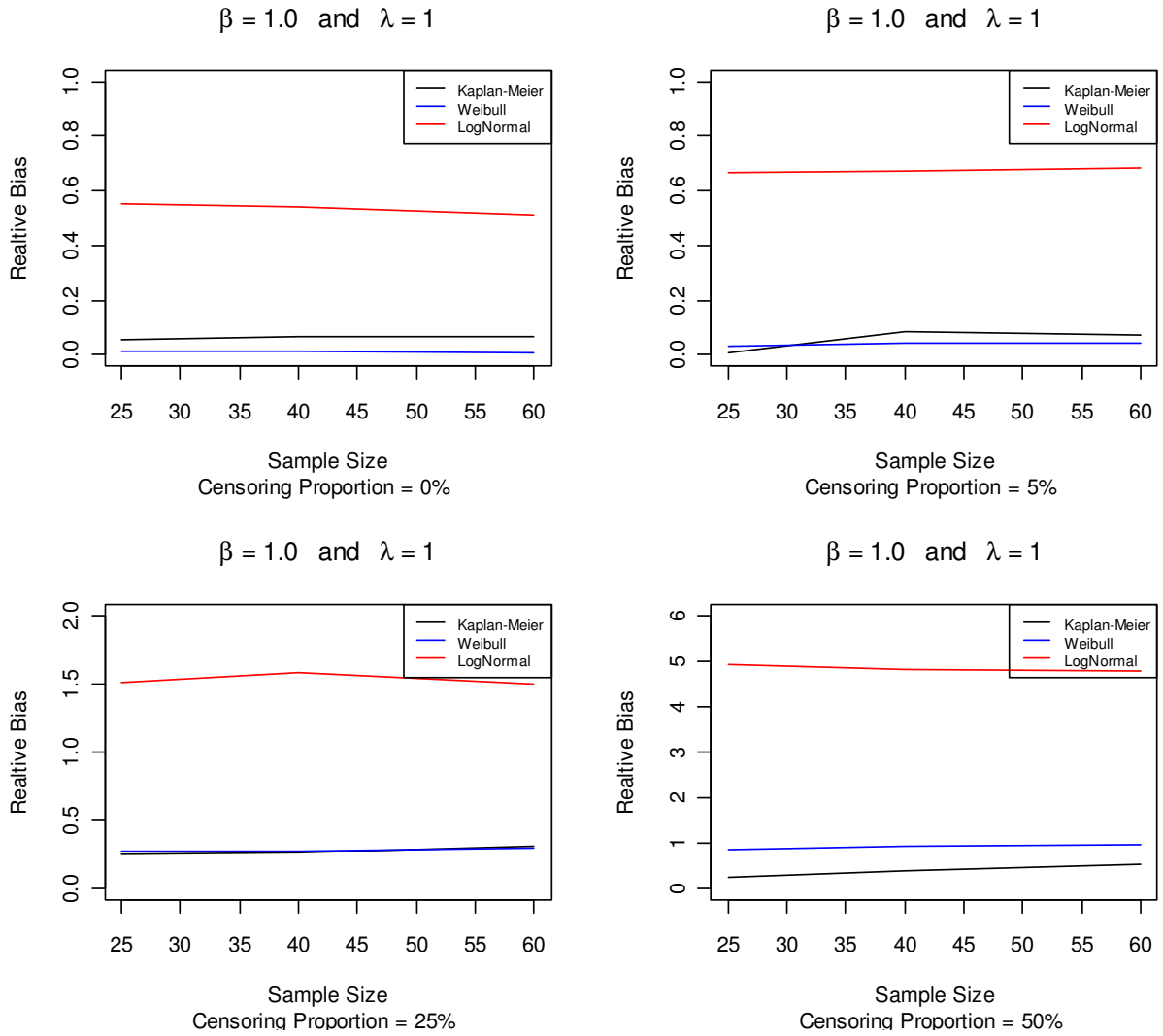
Figure 5.3 5^{th} Percentile Relative Bias Plots for Weibull Data with $\hat{a} = 1.0$ and $\tilde{e} = 1$



Overall, the relative bias estimates of the 5^{th} percentile in Figure 5.3 decrease with decreasing censoring rate. Relative Bias improves with increasing sample size from 25 to 40, but remain stable as the sample size increase from 40 to 60. For smaller censoring proportions and

smaller sample sizes, performance of the Kaplan-Meier estimator lies in between the other two. But for larger sample sizes, 40 and 60, it is the best irrespective of the censoring rate. The estimates are satisfactory here except for the highest censoring rate.

Figure 5.4 95^{th} Percentile Relative Bias Plots for Weibull Data with $\hat{a} = 1.0$ and $\tilde{e} = 1$



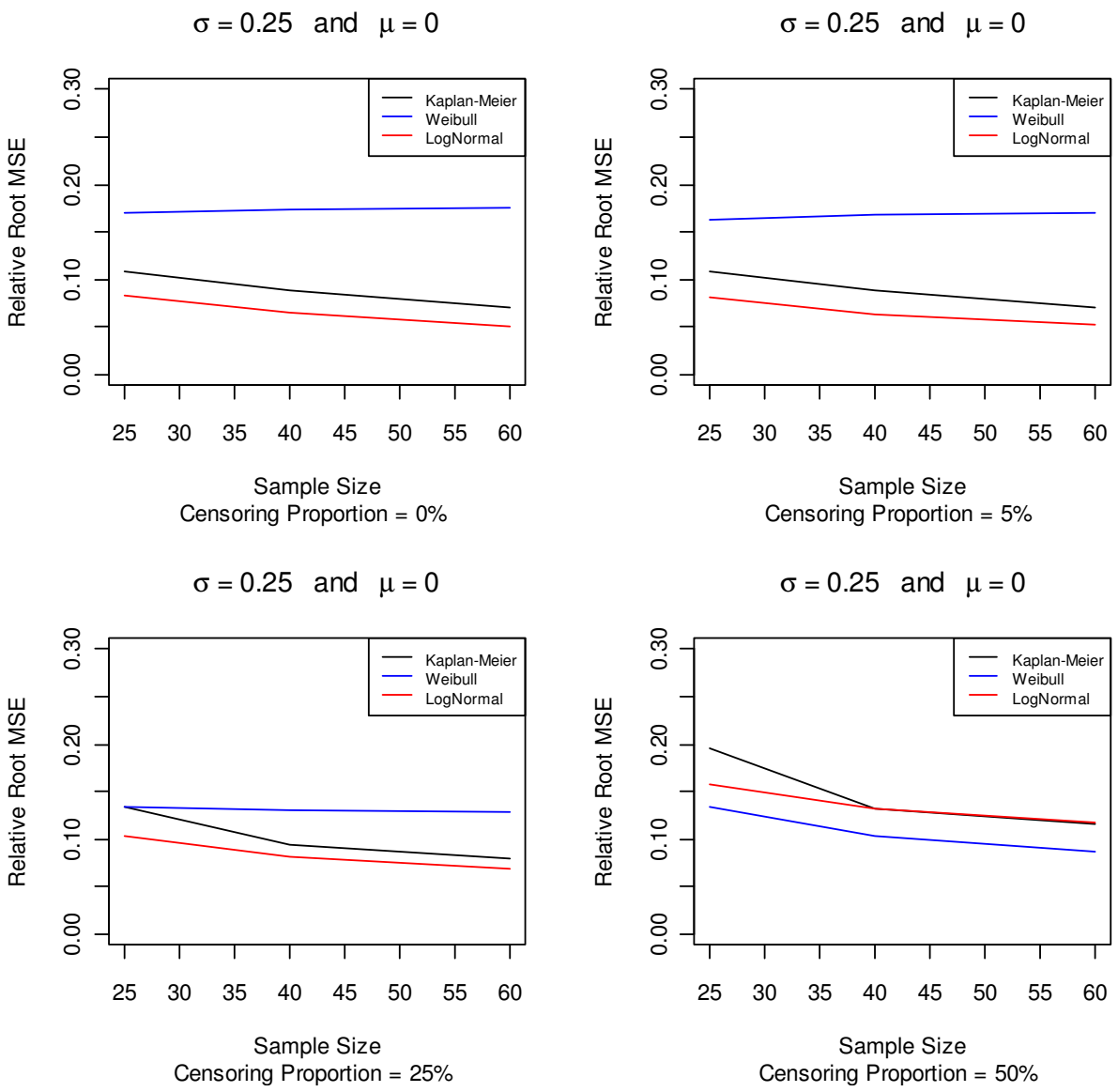
The vertical scales in the panels of Figure 5.4 are not the same in order to accommodate the vastly different relative bias that resulted from estimating the 95^{th} percentile of Weibull data. Here, estimates from the 'incorrect' lognormal analysis are much worse than the other two, whose profiles are almost identical. Overall, relative bias improves with decreasing censoring rate and is relatively stable across sample sizes. For larger censoring rates, Kaplan-Meier

estimates are even better than the estimates from the correct Weibull analysis. The lognormal estimates are clearly unsatisfactory.

For lower censoring rates the relative bias values are fairly small for both correctly estimated parametric estimates and non parametric estimates. This implies that they both are good estimates. With the increase of censoring rate, only the Kaplan-Meier estimates managed to have a smaller relative bias values.

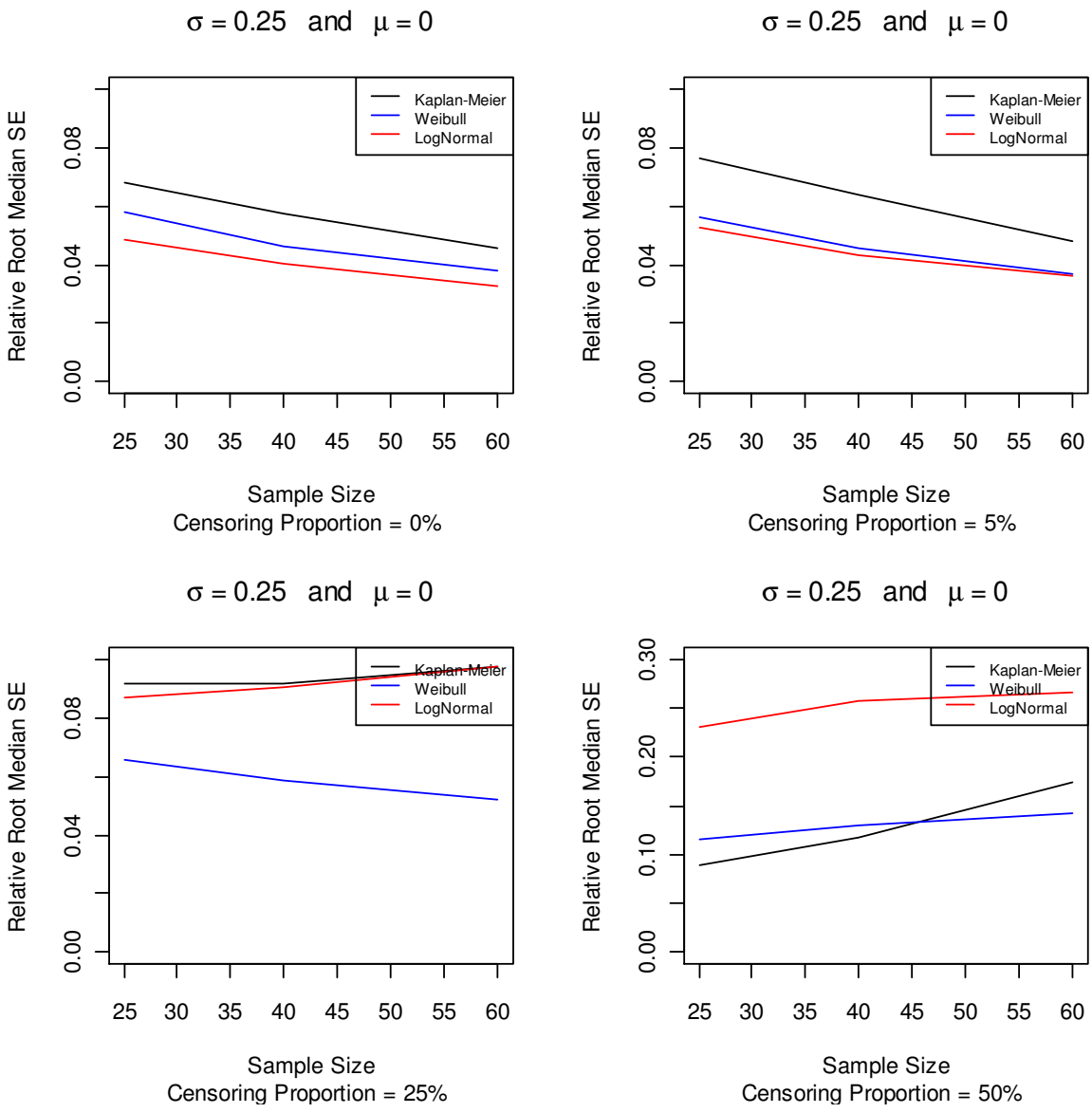
Figures 5.5 and 5.6 present plots of estimated relative mean square error vs sample size under various levels of censoring proportions for Lognormal data for the specified shape and location parameters. Here the blue profile represents the 'incorrect' Weibull estimates and the red profile the correct Lognormal estimates. The black profile represents the nonparametric Kaplan-Meier estimates.

Figure 5.5 **5^{th} Percentile Relative Root Mean Square Error Plots for Lognormal Data with $\sigma = 0.25$ and $\mu = 0$**



Relative root mean square errors of the 5th percentiles presented in Figure 5.5 based on lognormal likelihood and the Kaplan – Meier estimator improve with increasing sample size and decreasing censoring rate. Conversely, the 'incorrect' Weibull based estimates tend to perform better as the censoring rate increases. Further, for 50% censoring rate the 'incorrect' Weibull estimates tend to perform better than the correct lognormal estimates and all three estimates are very close to one another. Magnitudes of all the relative root mean square errors in Figure 5.5 are even below 0.2, which implies that these estimates are good.

Figure 5.6 95th Percentile Relative Root Median Square Error Plots for Lognormal Data with $\sigma = 0.25$ and $\mu = 0$

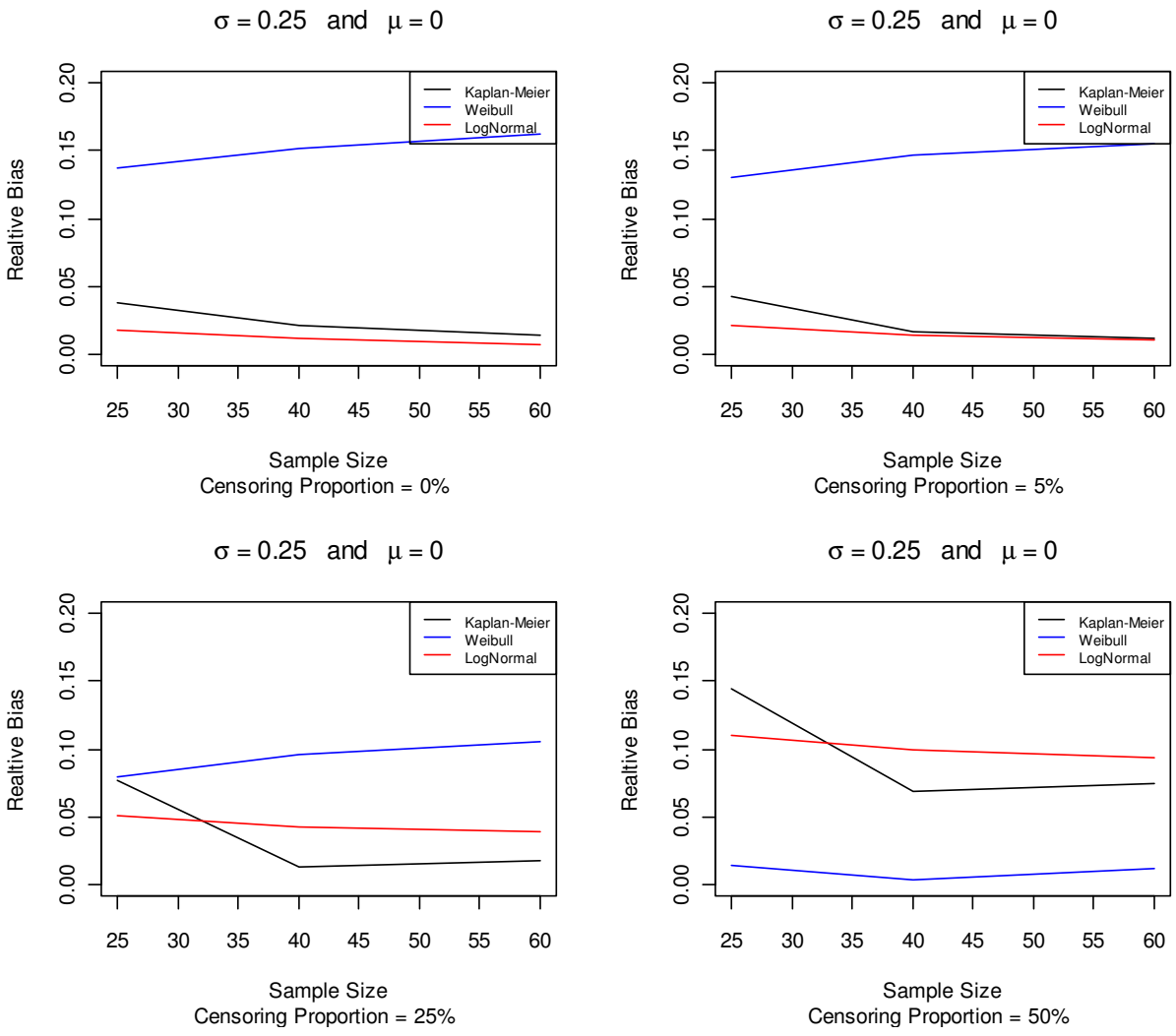


The vertical scales in the panels of Figure 5.6 are not the same in order to accommodate the vastly different relative root median square error that resulted from estimating the 95th percentile of lognormal data. Further, the MSE's are computed using the median of simulated square deviations instead of the mean because the distributions of these deviations are very highly right skewed. Here, estimates from the correct lognormal analysis are worse than the other two as the censoring rate increases. Overall, the relative root median square errors improve with decreasing censoring rate. For larger censoring proportions, Kaplan-Meier estimates of 95th percentiles are as good as the estimates from the 'incorrect' Weibull analysis.

Here, we have very small relative root mean square error values in all four plots in Figure 5.6, which implies that all the estimates are good.

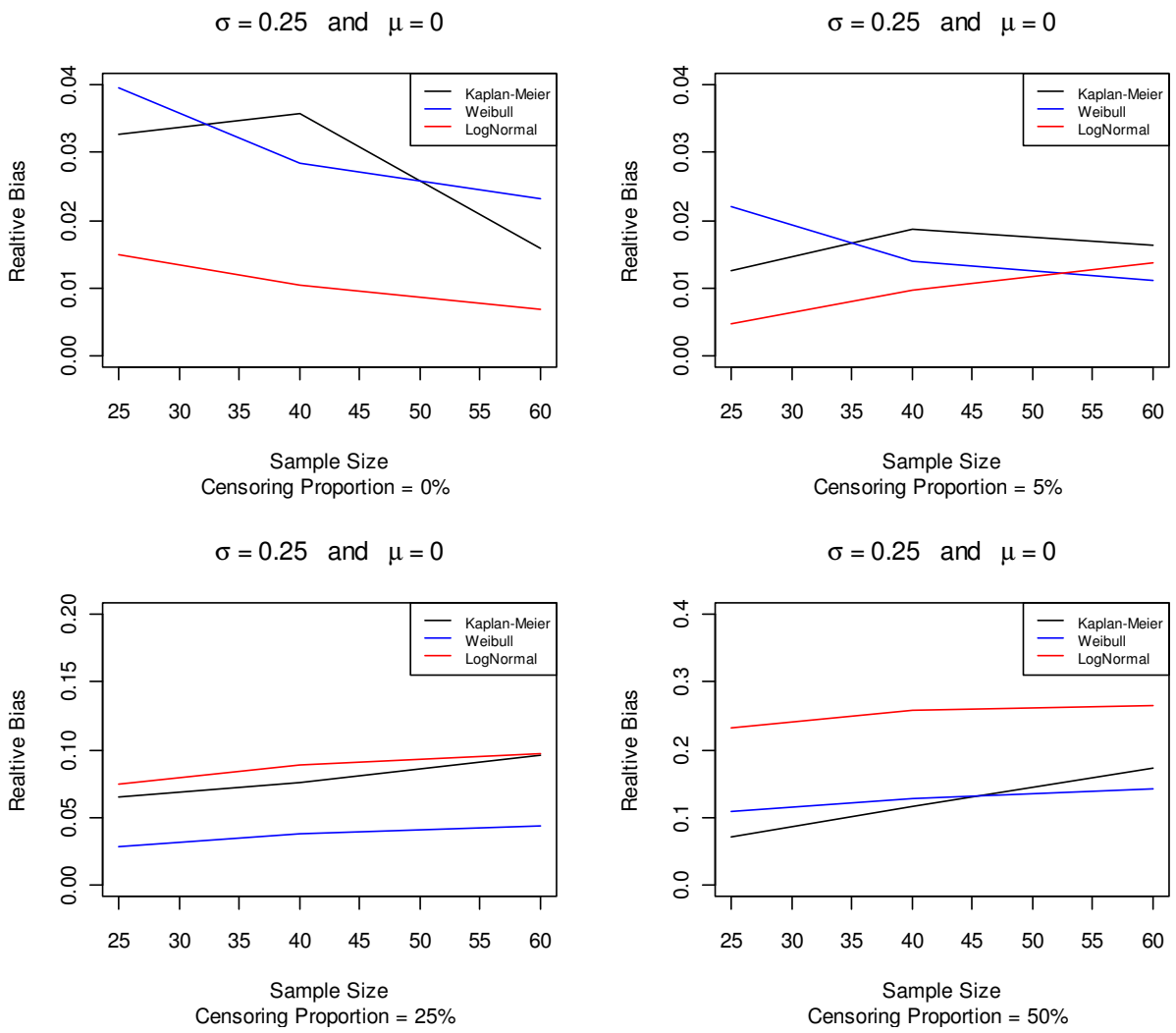
Figures 5.7 and 5.8 present plots of estimated relative bias vs sample size under various levels of censoring proportions for Lognormal data for the specified shape and location parameters. Here the blue profile represents the 'incorrect' Weibull estimates and the red profile the correct lognormal estimates. The black profile represents the nonparametric Kaplan-Meier estimates.

Figure 5.7 **5th Percentile Relative Bias Plots for Lognormal Data with $\sigma = 0.25$ and $\mu = 0$**



For larger sample sizes (40 and 60) performance of Lognormal estimates and Kaplan-Meier estimates stay closer to each other. The interesting feature here is the behavior of the 'incorrect' Weibull estimates, which improves with the increase of censoring rate and ultimately tends to be the best estimates. All of the above estimates are satisfactory.

Figure 5.8 **95th Percentile Relative Bias Plots for Lognormal Data with $\sigma = 0.25$ and $\mu = 0$**



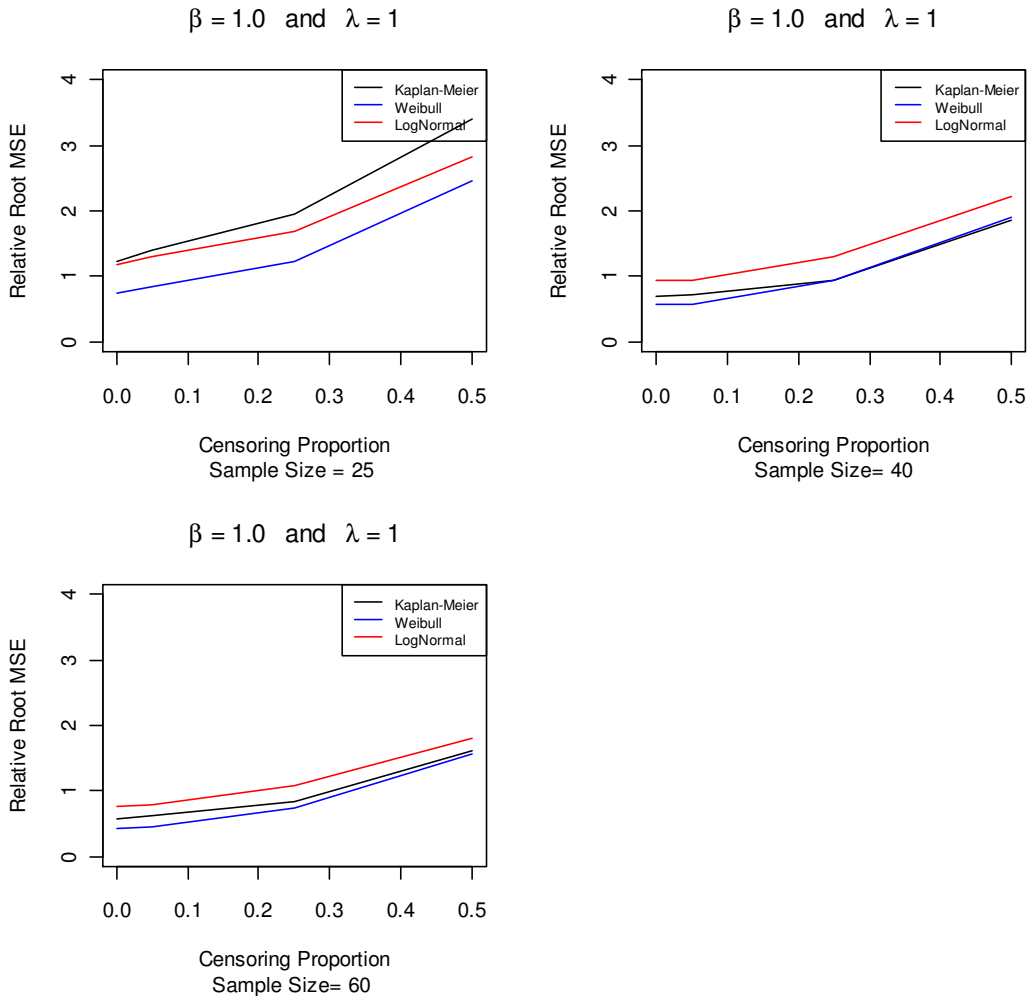
The vertical scales in the panels of Figure 5.8 are not the same in order to accommodate the vastly different relative biases that resulted from estimating the 95th percentile of lognormal data. Here again the both the other estimates perform better than the correct lognormal estimates as the censoring rate increase.

Here I presented only the plots for one set of parameters from each distribution. The other plots can be found in Appendix B.

Relative Root Mean/Median Square Error vs Censoring Proportion Plots

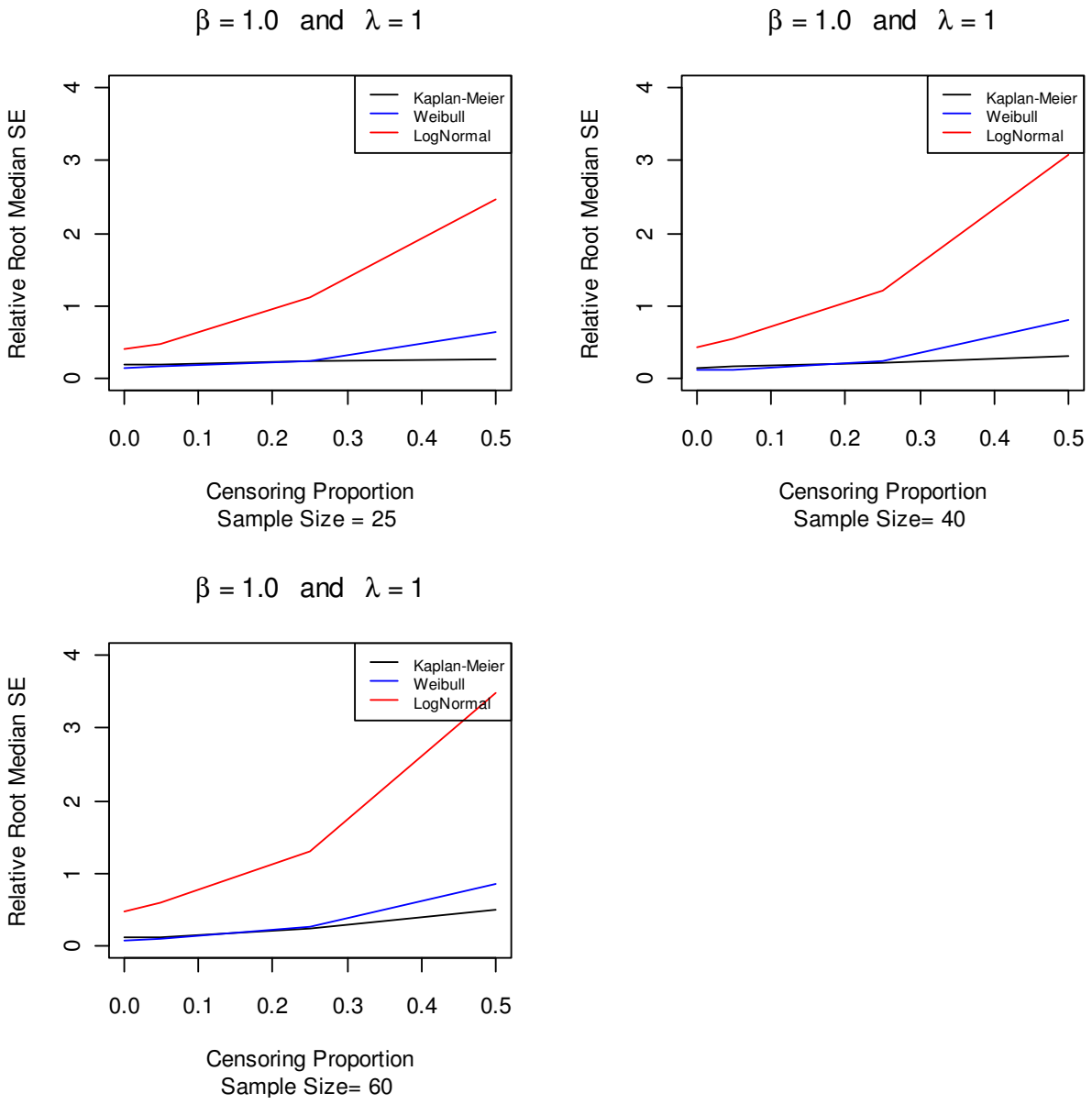
Figures 5.9 and 5.10 present plots of estimated relative root mean square error vs censoring proportion under various sample sizes for Weibull data for the specified shape and scale parameters. Here the blue profile represents the correct Weibull estimates and the red profile the 'incorrect' lognormal estimates. The black profile represents the nonparametric Kaplan-Meier estimates.

Figure 5.9 **5^{th} Percentile Relative Root Mean Square Error vs Censoring Proportion Plots for Weibull Data with $\hat{a} = 1.0$ and $\ddot{e} = 1$**



Estimates of the 5^{th} percentile in Figure 5.9 improve with decreasing censoring rate and increasing sample size.

Figure 5.10 **95th Percentile Relative Root Median Square Error vs Censoring Proportion Plots for Weibull Data with $\hat{a} = 1.0$ and $\ddot{e} = 1$**



Again, the estimates of the 95th percentile in Figure 5.10 improve with decreasing censoring rate.

Regression and Model Fitting

To aid in studying the relationship of the variables of interest (sample size, censoring proportion, parameter values, and the method used) to the quantile estimates, we tried fitting several regression models with these values as the independent variables and relative root mean (or median) square error as the response.

Regression for Weibull Data Estimates for 5th Percentile

When considering the Weibull data, the following are the variables of interest.

X1 = Sample Size - n
X2 = Scale Parameter - Lambda
X3 = Shape Parameter - Beta
X4 = Censoring Proportion - Phi
X5 = Quantile Considered - Zetap

X6 = 1 for KM 0 - o/w
X7 = 1 for LN 0 - o/w } Baseline is Weibull

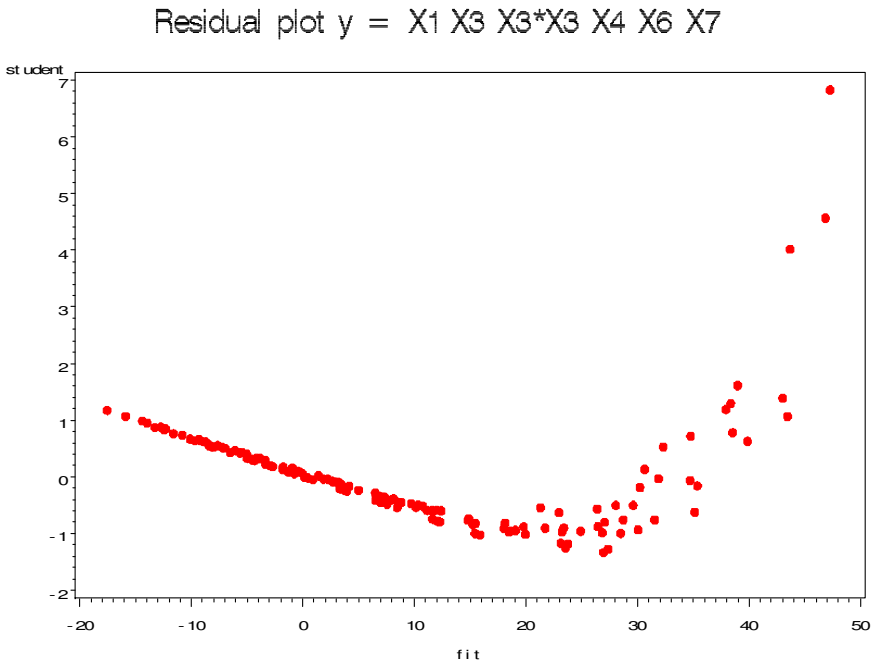
X8 = Squared Shape Parameter - X3*X3 and

Y = Relative Root Mean Square Error

Table 5.1 Results of Model Fitting for Weibull Data Estimates of 5 th Percentile

Model	R-Squared	Terms Significant
Y = X1 X3 X3*X3 X4 X6 X7	0.514577	X1, X3, X3*X3 and X4
Y = X1 X1*X1 X3 X3*X3 X4 X6 X7	0.520888	X1, X3, X3*X3 and X4

Figure 5.11 Residual Plot for Weibull Data Estimates of 5th Percentile



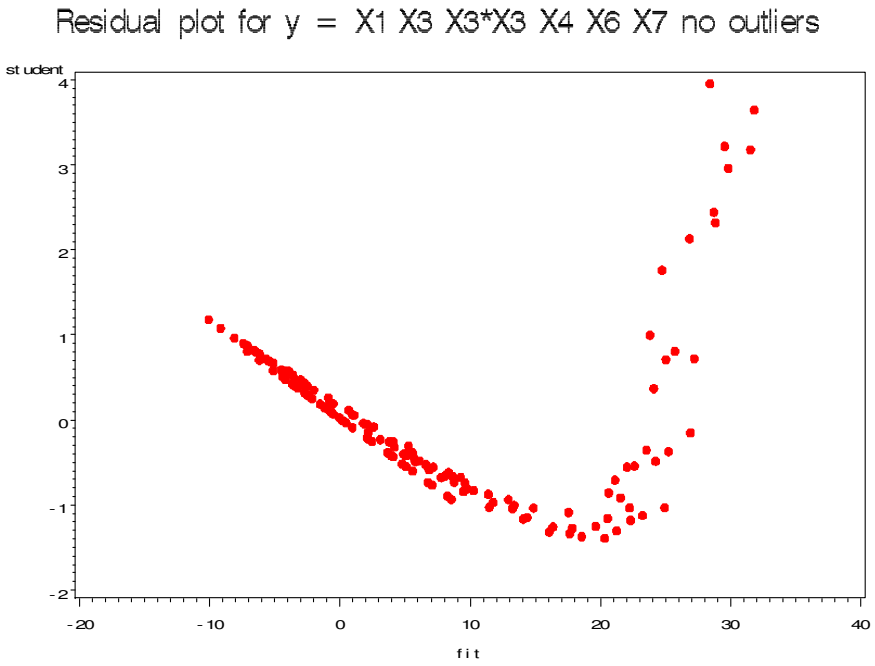
The R-Squared values in Table 5.1 are just above 0.5. It can be seen from Figure 5.11 that there are three possible outliers in the studentized residual plot. We then tried to refit these models after removing these three outliers.

Table 5.2 Results of Model Fitting for Weibull Data Estimates of 5 th Percentile Without Outliers

Model	R-Squared	Terms Significant
Y = X1 X3 X3*X3 X4 X6 X7	0.607403	X3, X3*X3 and X4
Y = X1 X1*X1 X3 X3*X3 X4 X6 X7	0.607622	X3, X3*X3 and X4

Since there is not much difference in R-Squared values of the two models, we will proceed with the first model, which has fewer independent variables.

Figure 5.12 Residual Plot Without Outliers for Weibull Data Estimates of 5th Percentile

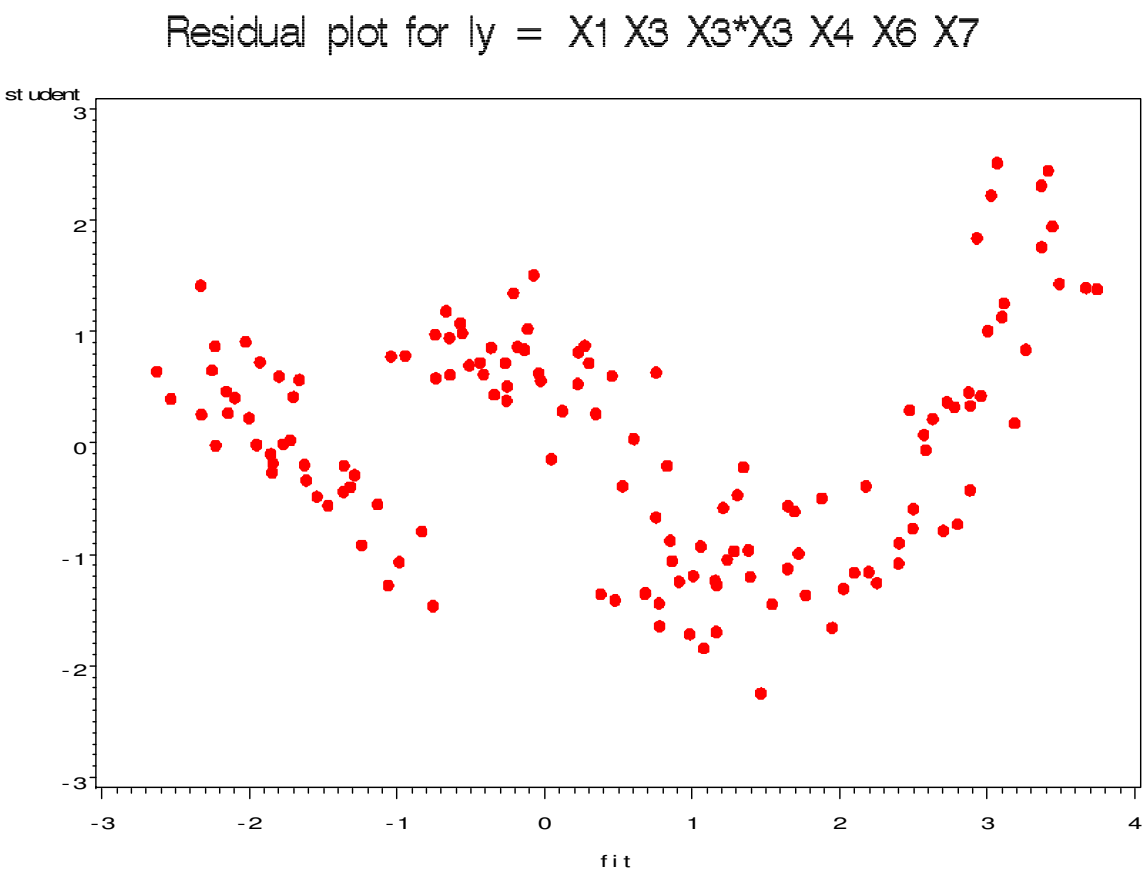


Since the residual plot in Figure 5.12 is not satisfactory, we tried fitting the first model in Table 5.2 with response equal to log(Y).

Table 5.3 Results of Model Fitting for Weibull Data Estimates of 5 th Percentile Without Outliers using log(Y) as the Response

Model	R-Squared	Terms Significant
Log(Y) = X1 X3 X3*X3 X4 X6 X7	0.971269	X1, X3, X3*X3, X4, X6 and X7

Figure 5.13 Residual Plot Without Outliers for Weibull Data Estimates of 5th Percentile using log(Y) as the Response



Both the residual plot and the large value of the coefficient of determination indicate an adequate fit.

Parameter Estimates for the Final Model

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	5.029690160	0.13219641	38.05	<.0001
X1	-0.015164383	0.00183786	-8.25	<.0001
X3	-4.334685978	0.11305432	-38.34	<.0001
X3*X3	0.597065737	0.01925638	31.01	<.0001
X4	1.933864969	0.13538847	14.28	<.0001
X6	0.299313023	0.06417343	4.66	<.0001
X7	0.374719065	0.06417343	5.84	<.0001

According to the parameters estimated for this model:

All the terms (Sample size-X1, Shape parameter -X3, Squared shape parameter -X3*X3, Censoring proportion-X4, dummy variables–X6 and X7 to represent the method used) are statistically significant. From the regression output above, leaving all other variables fixed, we we estimate that the logarithm of relative root mean square error in estimating the 5^{th} percentile of Weibull data : (i) decreases by 0.015 per unit increase in sample size; (ii) increase by 1.933 per unit increase in censoring rate; (iii) is quadratic in the shape parameter, first decreasing and then increasing and (iv) is larger than the correct Weibull analysis when either the Kaplan-Meier estimator or incorrect lognormal analysis used.

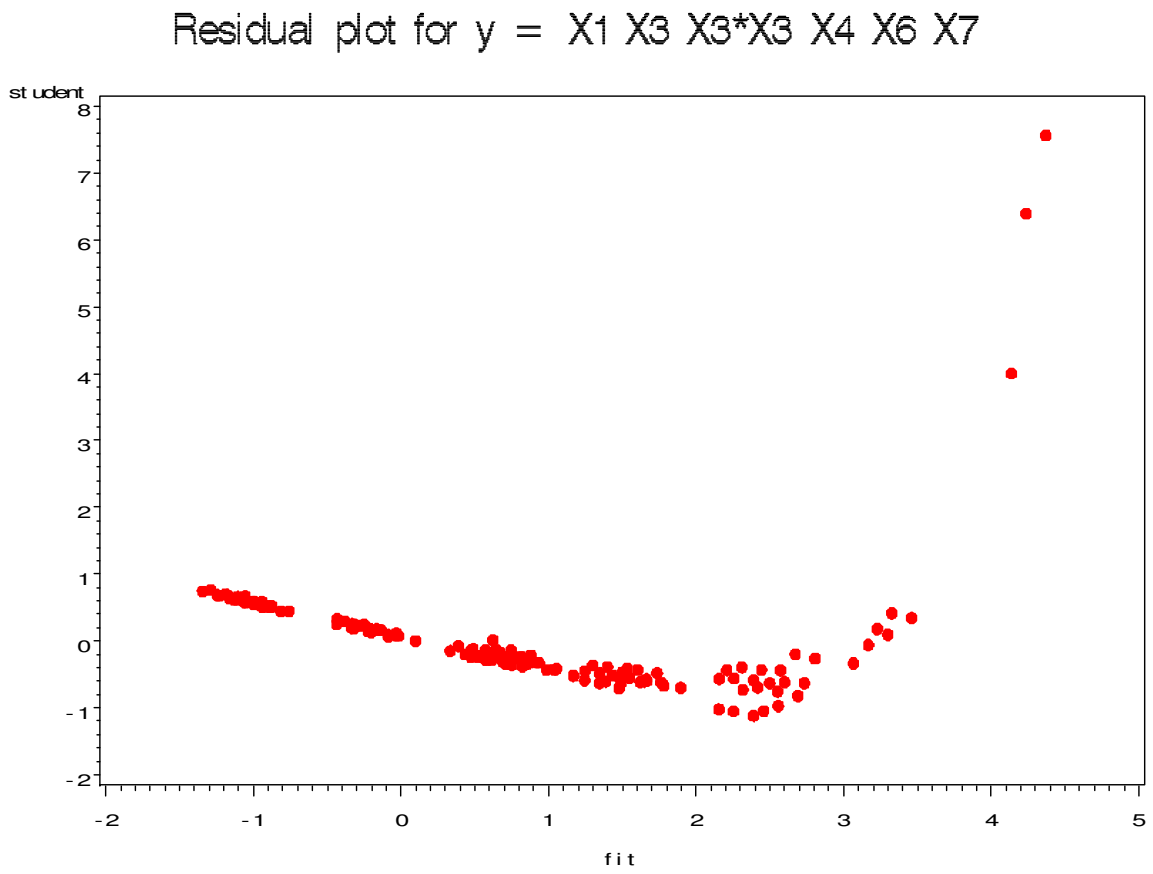
Next I fitted regression model to estimates of the 95^{th} percentile of Weibull data with y = relative root median square error as the response.

Regression for Weibull Data Estimates for 95^{th} Percentile

Table 5.4 Results of Model Fitting for Weibull Data Estimates of 95^{th} Percentile

Model	R-Squared	Terms Significant
Y = X1 X3 X3*X3 X4 X6 X7	0.338030	X3, X4, and X7
Y = X1 X1*X1 X3 X3*X3 X4 X6 X7	0.338130	X4 and X7

Figure 5.14 Residual Plot for Weibull Data Estimates of 95th Percentile



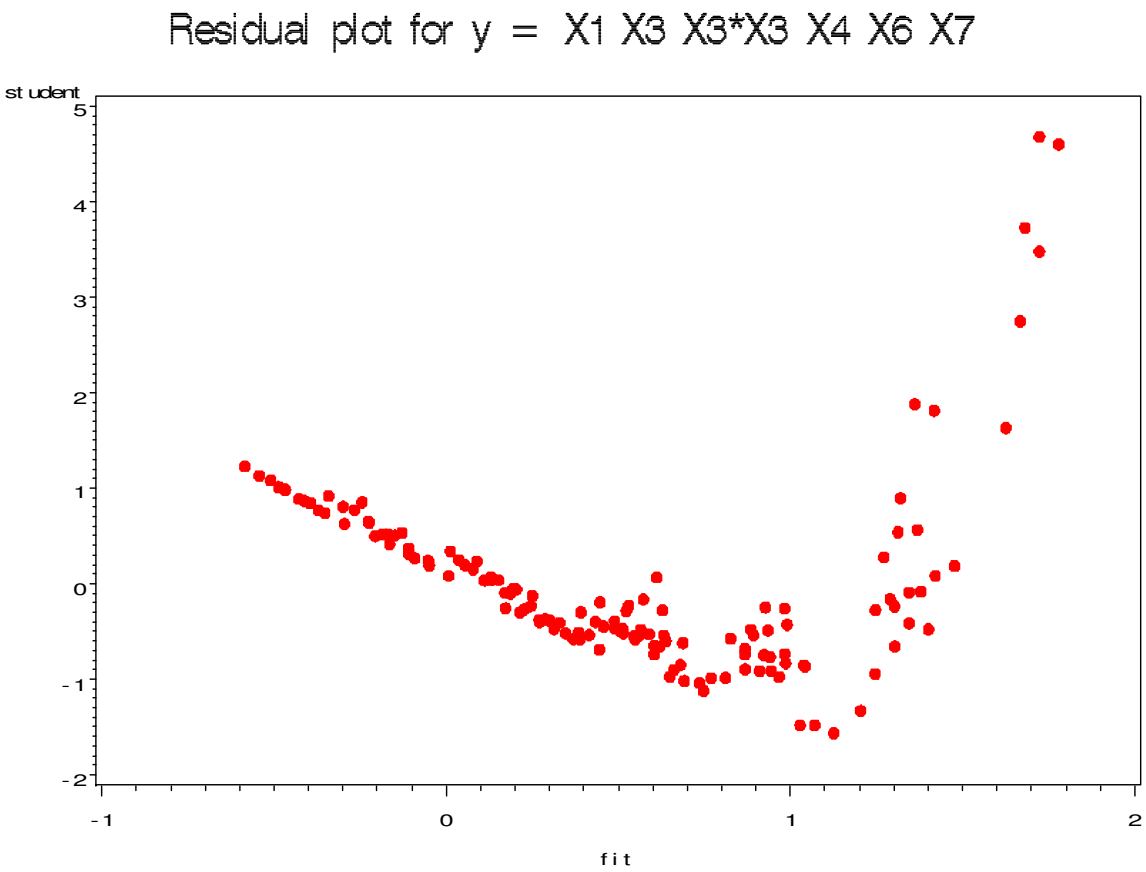
The R-Squared values in Table 5.4 are even smaller than 0.5. It can be seen from Figure 5.14 that there are three possible outliers in the residual plot. We then tried to refit these models after removing these three outliers.

Table 5.5 Results of Model Fitting for Weibull Data Estimates of 95th Percentile Without Outliers

Model	R-Squared	Terms Significant
Y = X1 X3 X3*X3 X4 X6 X7	0.569883	X3, X4 and X7
Y = X1 X1*X1 X3 X3*X3 X4 X6 X7	0.569961	X3, X3*X3, X4, X6 and X7

Since there is not much difference in R-Squared values of the two models, we will proceed with the first model, which has fewer independent variables.

Figure 5.15 Residual Plot Without Outliers for Weibull Data Estimates of 95th Percentile

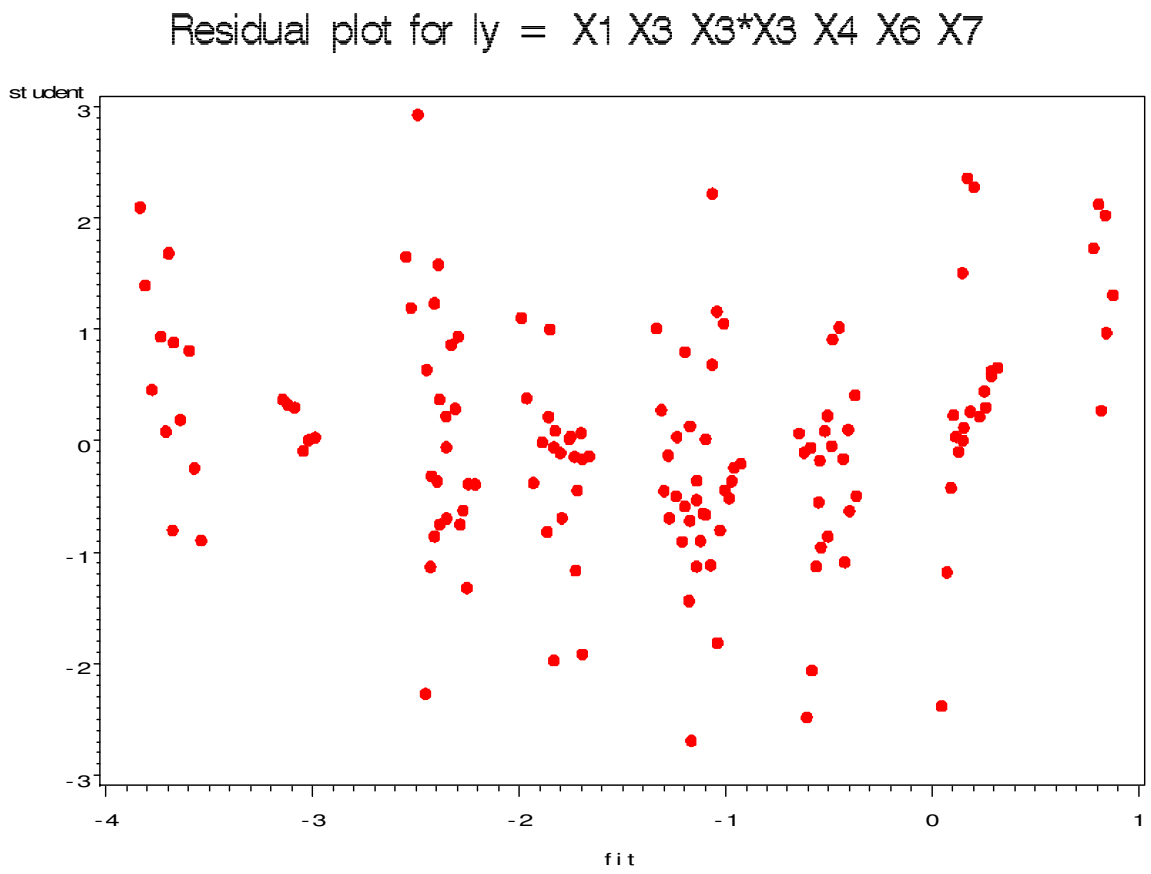


Since the residual plot in Figure 5.15 is not satisfactory, we tried fitting the first model in Table 5.5 with response equal to log(Y).

Table 5.6 Results of Model Fitting for Weibull Data Estimates of 95th Percentile Without Outliers using log(Y) as the Response

Model	R-Squared	Terms Significant
Log(Y) = X1 X3 X3*X3 X4 X6 X7	0.946355	X3, X3*X3, X4, X6 and X7

Figure 5.16 Residual Plot Without Outliers for Weibull Data Estimates of 95th Percentile using log(Y) as the Response



Both the residual plot and the large value of the coefficient of determination indicate an adequate fit.

Parameter Estimates for the Final Model

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.798241396	0.12292614	6.49	<.0001
X1	0.001631012	0.00171059	0.95	0.3421
X3	-1.587568786	0.10587304	-15.00	<.0001
X3*X3	0.187667438	0.01803230	10.41	<.0001
X4	2.762400999	0.12679489	21.79	<.0001
X6	-1.425715983	0.06054399	-23.55	<.0001
X7	-1.325637864	0.06054399	-21.90	<.0001

All the terms (Shape parameter-X3, Squared shape parameter-X3*X3, Censoring proportion-X4, and the two dummy variables-X6 and X7 to represent the method used) are statistically significant except the Sample size-X1. From the regression output above, leaving all other variables fixed, we estimate that the logarithm of relative root median square error estimating the 95^{th} percentile of Weibull data: (i) increases by 2.762 per unit increase in censoring rate; (ii) is quadratic in the shape parameter, first decreasing and then increasing and (iii) is less than the correct Weibull analysis when either the Kaplan-Meier estimator or incorrect lognormal analysis is used. Somewhat surprisingly, (iii) is different from what is seen in Figure 5.10.

Regression for Lognormal Data Estimates for 5th Percentile

When considering the Lognormal data, the following are the variables of interest.

X1 = Sample Size - n

X2 = Mu

X3 = Sigma

X4 = Censoring Proportion - Phi

X5 = Quantile Considered - Zetap

X6 = 1 for KM 0 - o/w

X7 = 1 for W 0 - o/w } Baseline is Lognormal

X8 = Squared Shape Parameter - X3*X3

Y = Relative Root Mean Square Error

Note that since one set of parameters for lognormal data (Mu=0, Sigma =10) correspond to a distribution that was so skewed as to be not realistic, it was dropped from the regression analysis.

Table 5.7 Results of Model Fitting for Lognormal Data Estimates of 5th Percentile

Model	R-Squared	Terms Significant
Y = X1 X3 X3*X3 X4 X6 X7	0.666915	X1 and X4
Y = X1 X1*X1 X3 X3*X3 X4 X6 X7	0.676019	X4

The R-Squared values in Table 5.7 are around 0.6. It can be seen from Figure 5.17 that there are two possible outliers in the residual plot. We then tried to refit these models after removing these two outliers.

Figure 5.17 Residual Plot for Lognormal Data Estimates of 5th Percentile

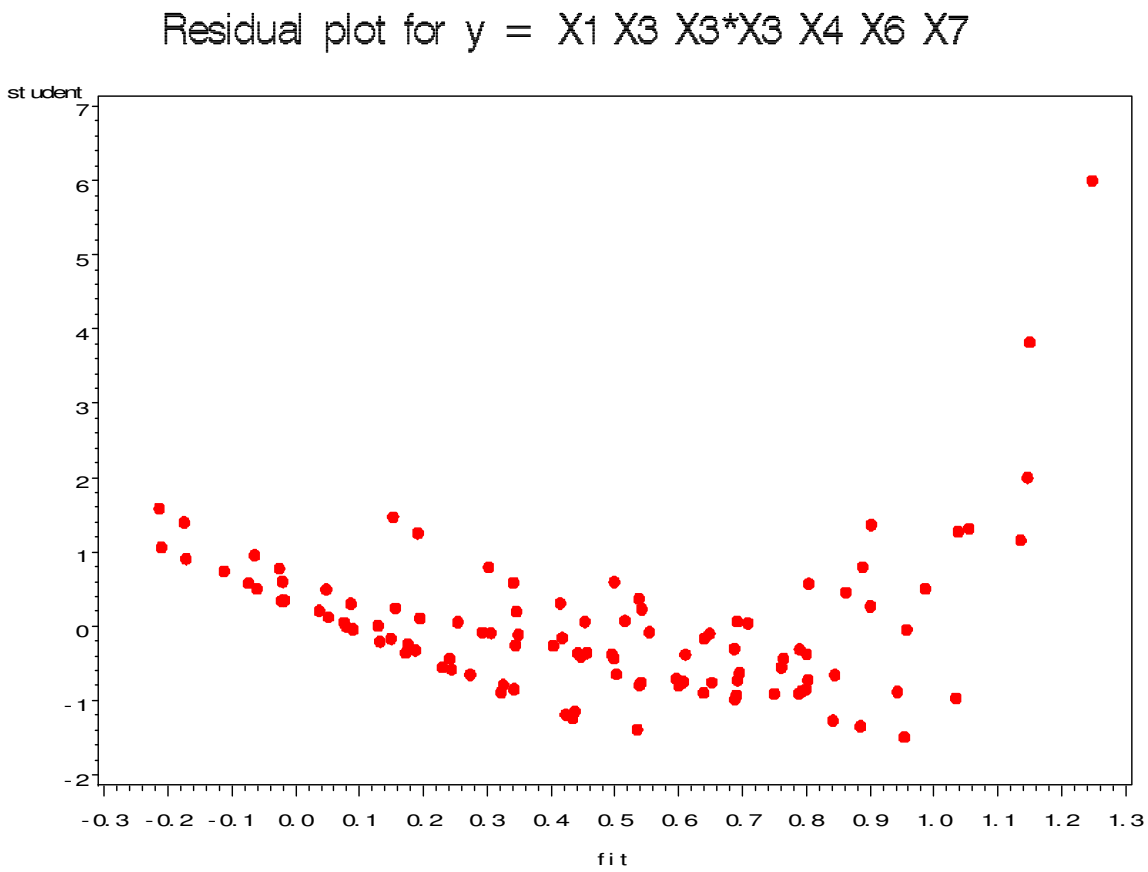
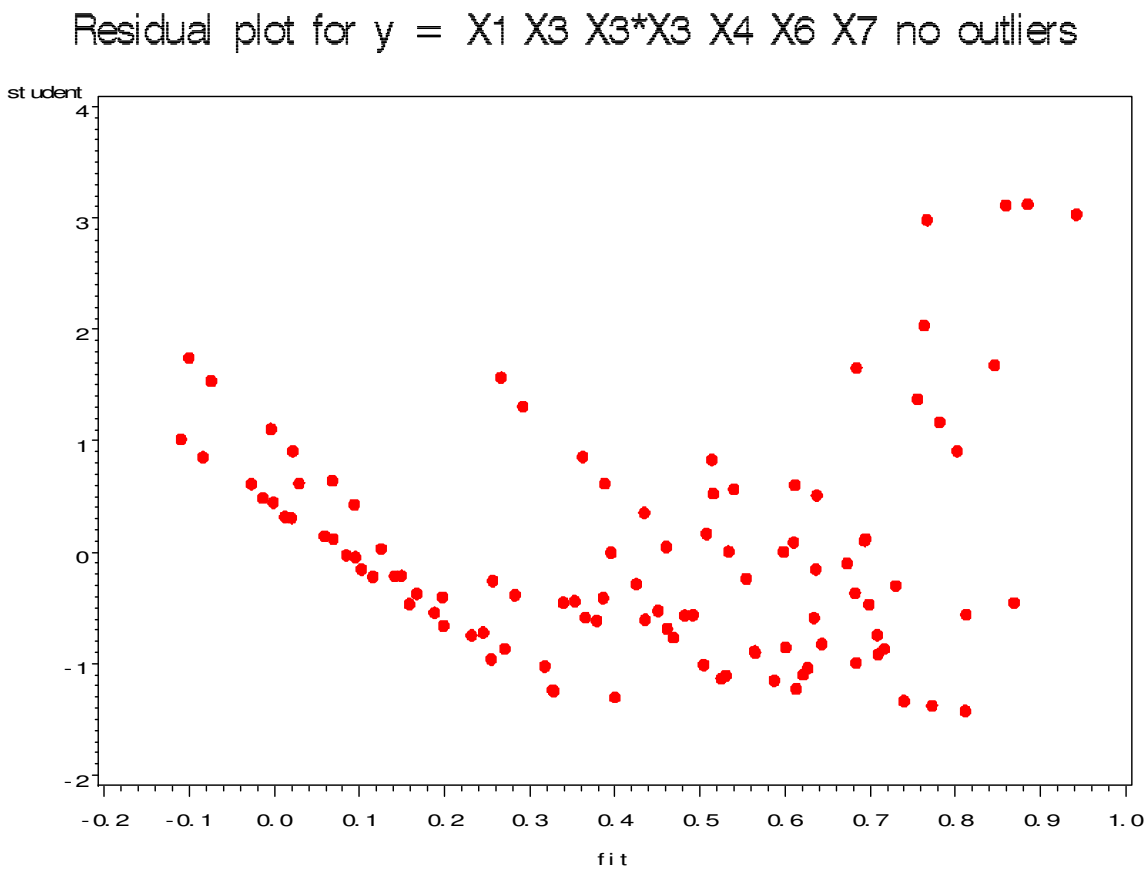


Table 5.8 Results of Model Fitting for Lognormal Data Estimates of 5th Percentile Without Outliers

Model	R-Squared	Terms Significant
Y = X1 X3 X3*X3 X4 X6 X7	0.747084	X1 and X4
Y = X1 X1*X1 X3 X3*X3 X4 X6 X7	0.749236	X4

Since there is not much difference in R-Squared value of the two models, we will proceed with the first model, which has fewer independent variables.

Figure 5.18 Residual Plot Without Outliers for Lognormal Data Estimates of 5th Percentile

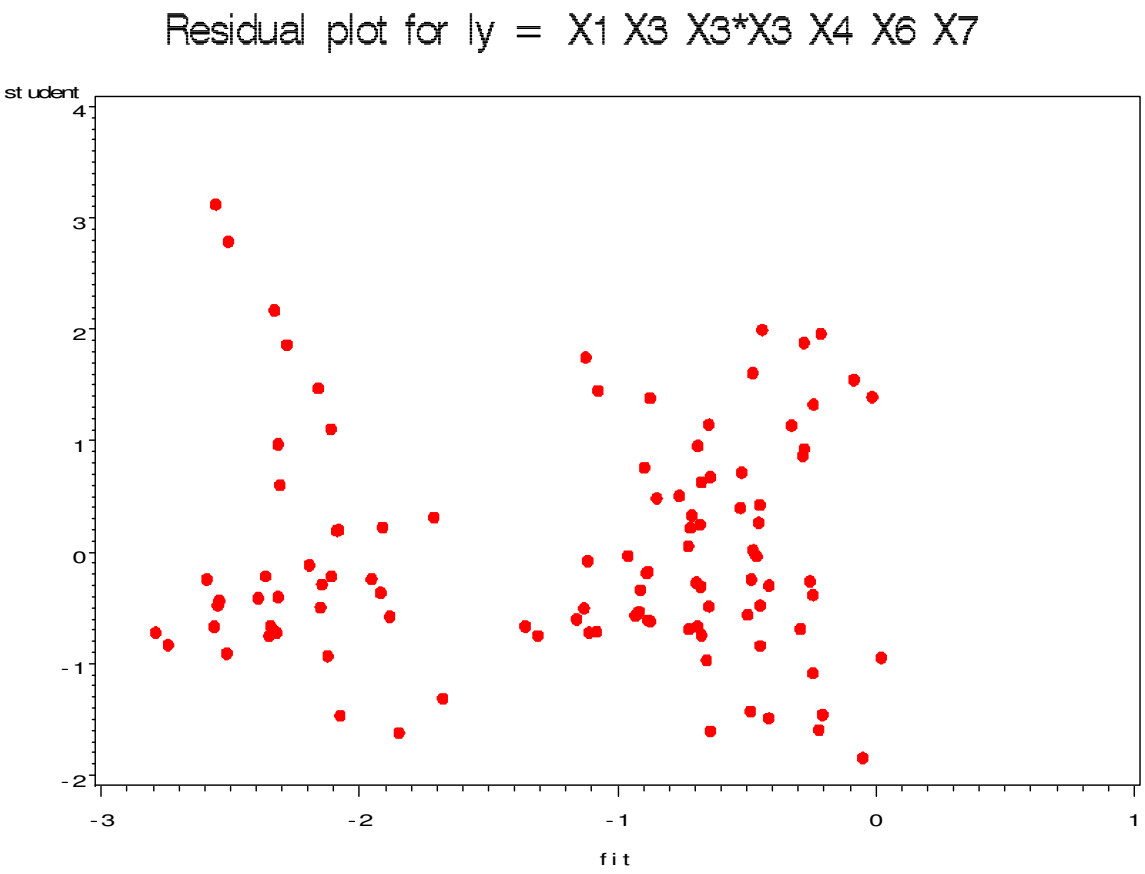


Since the residual plot in Figure 5.18 is not satisfactory, we tried fitting the first model in Table 5.8 with response equal to log(Y).

Table 5.9 Results of Model Fitting for Lognormal Data Estimates of 5th Percentile Without Outliers using log(Y) as the Response

Model	R-Squared	Terms Significant
Log(Y) = X1 X3 X3*X3 X4 X6 X7	0.908286	X1, X3, X3*X3, X4, X6 and X7

Figure 5.19 Residual Plot Without Outliers for Lognormal Data Estimates of 5th Percentile using log(Y) as the Response



Both the residual plot and the large value of the coefficient of determination indicate an adequate fit.

Parameter Estimates for the final model

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-2.792657046	0.12555517	-22.24	<.0001
X1	-0.011357380	0.00186728	-6.08	<.0001
X3	2.950262019	0.26161433	11.28	<.0001
X3*X3	-0.832186563	0.15084846	-5.52	<.0001
X4	0.964525272	0.13813576	6.98	<.0001
X6	0.198419222	0.06502557	3.05	0.0029
X7	0.233212552	0.06502557	3.59	0.0005

All the terms (Sample size-X1, Shape parameter-X3, Squared shape parameter-X3*X3, Censoring proportion-X4), the two dummy variables-X6 and X7 to represent the method used are statistically significant. From the regression output above, leaving all other variables fixed, we estimate that the logarithm of relative root mean square error in estimating the 5^{th} percentile of lognormal data : (i) decreases by 0.011 per unit increase in sample size; (ii) increases by 0.965 per unit increase in censoring rate; (iii) is quadratic in sigma, first increasing and then decreasing and (iv) is larger than correct lognormal analysis when either the Kaplan-Meier estimator or incorrect Weibull analysis is used.

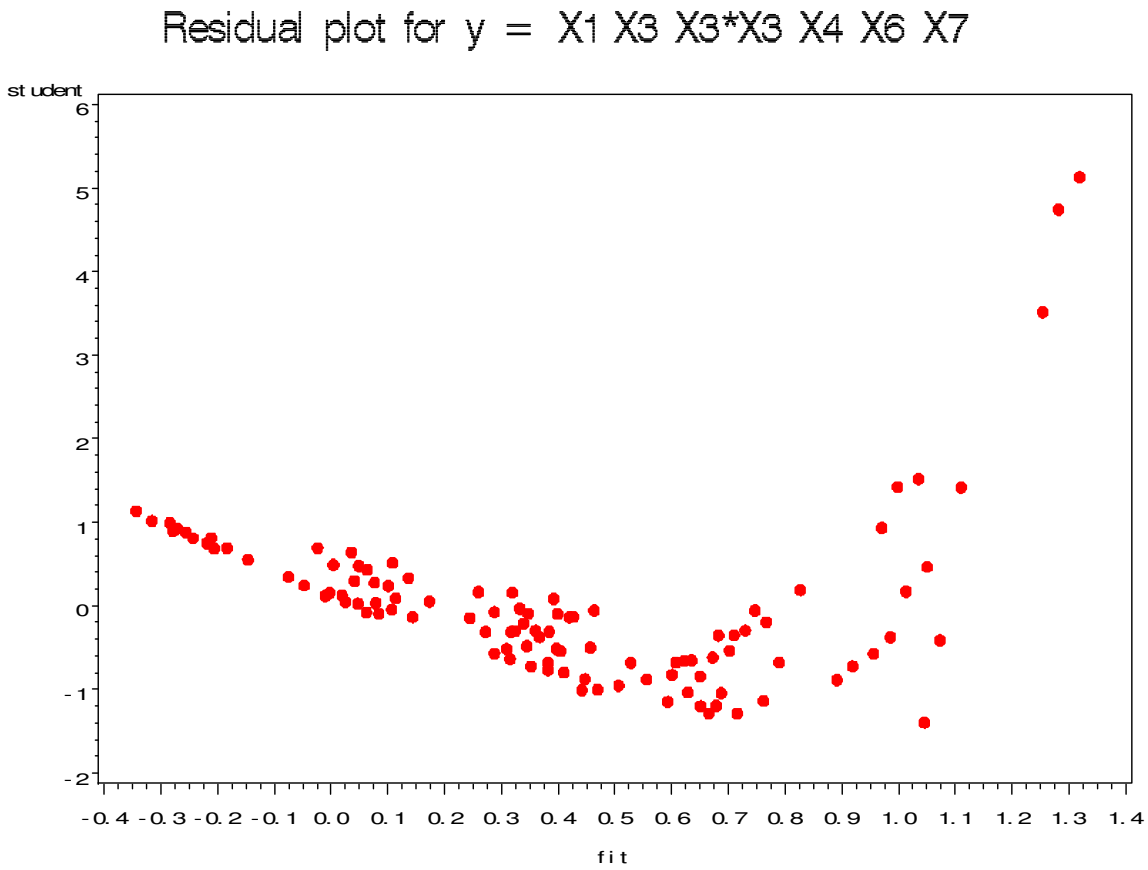
Next I fitted a regression model to estimates of the 95^{th} percentile of lognormal data with y= relative root median square error as the response.

Regression for Lognormal Data Estimates for 95^{th} Percentile

Table 5.10 Results of Model Fitting for Lognormal Data Estimates of 95^{th} Percentile

Model	R-Squared	Terms Significant
Y = X1 X3 X3*X3 X4 X6 X7	0.556041	X4
Y = X1 X1*X1 X3 X3*X3 X4 X6 X7	0.556046	X4

Figure 5.20 Residual Plot for Lognormal Data Estimates of 95th Percentile



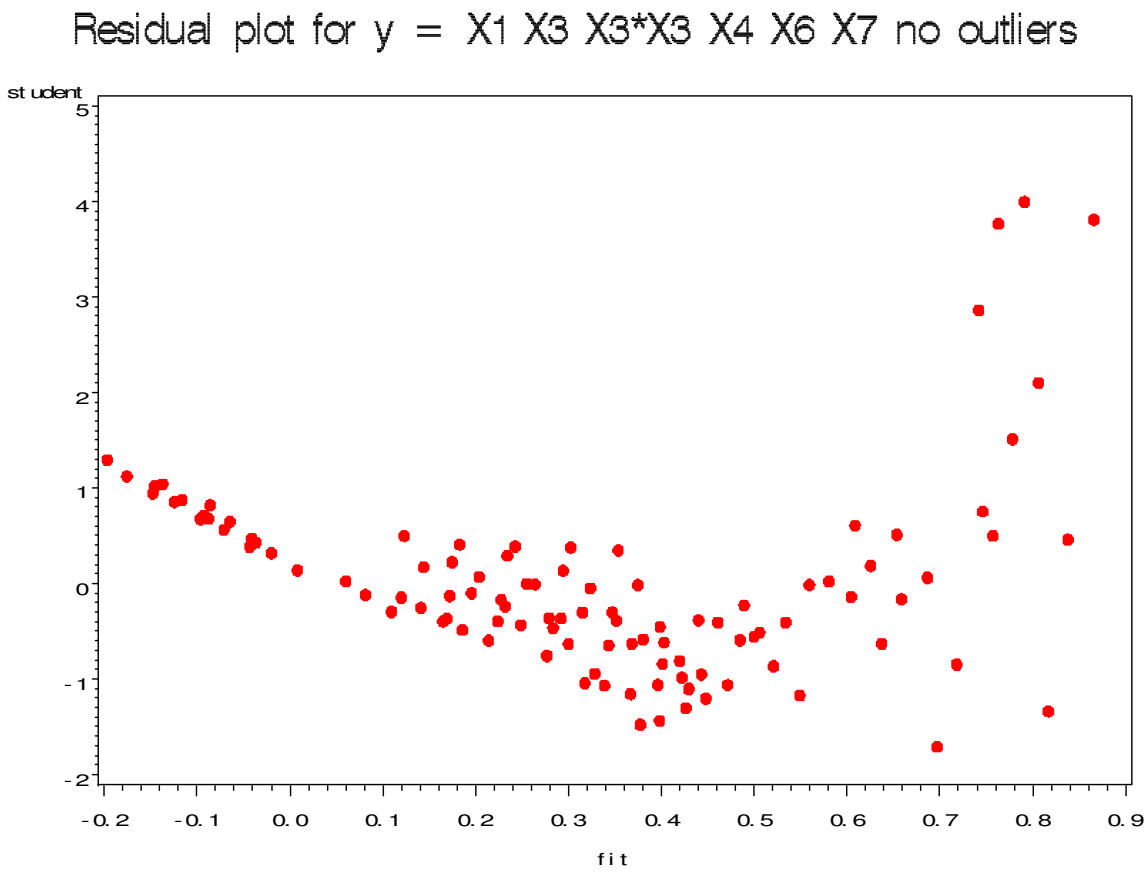
The R-Squared values in Table 5.10 are just above 0.5. It can be seen from Figure 5.20 that there are three possible outliers in the residual plot. We then tried to refit these models after removing these three outliers.

Table 5.11 Results of Model Fitting for Lognormal Data Estimates of 95th Percentile Without Outliers

Model	R-Squared	Terms Significant
Y = X1 X3 X3*X3 X4 X6 X7	0.648109	X4
Y = X1 X1*X1 X3 X3*X3 X4 X6 X7	0.648128	X4

Since there is not much difference in R-Squared value of the two models, we will proceed with the first model, which has fewer independent variables.

Figure 5.21 Residual Plot Without Outliers for Lognormal Data Estimates of 95th Percentile

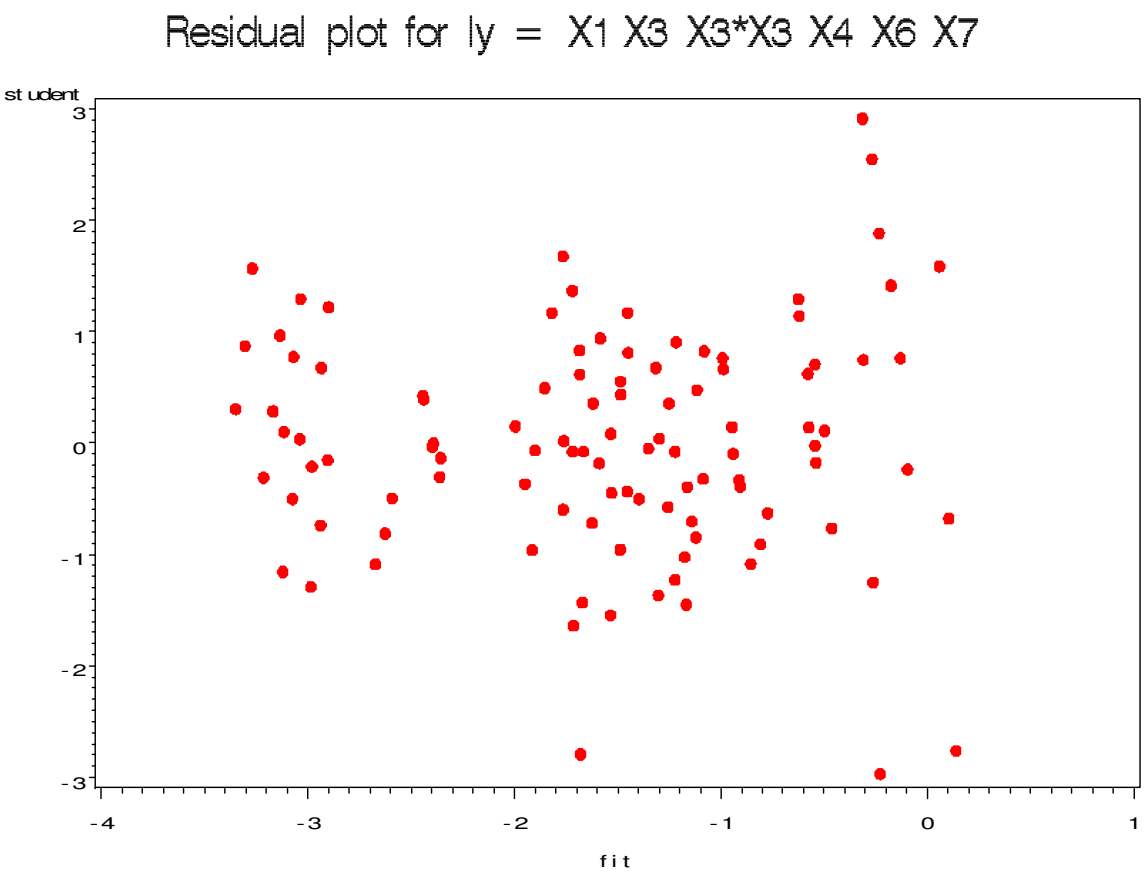


Since the residual plot in Figure 5.21 is not satisfactory, we tried fitting the first model in Table 5.11 with response equal to log(Y).

Table 5.12 Results of Model Fitting for Lognormal Data Estimates of 95th Percentile Without Outliers using log(Y) as the Response

Model	R-Squared	Terms Significant
Log(Y) = X1 X3 X3*X3 X4 X6 X7	0.924570	X3, X3*X3, X4 and X7

Figure 5.22 Residual Plot Without Outliers for Lognormal Data Estimates of 95th Percentile using log(Y) as the Response



Both the residual plot and the large value of the coefficient of determination indicate an adequate fit.

Parameter Estimates for the Final Model

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-3.703388591	0.12665948	-29.24	<.0001
X1	-0.002317387	0.00186924	-1.24	0.2180
X3	3.132362528	0.26414595	11.86	<.0001
X3*X3	-0.959118774	0.15232778	-6.30	<.0001
X4	2.710963235	0.13956880	19.42	<.0001
X6	0.004832928	0.06639877	0.07	0.9421
X7	-0.229553254	0.06639877	-3.46	0.0008

Only the terms Sigma-X3, Sigma squared-X3*X3, and Censoring proportion-X4 and dummy variable for Weibull analysis are statistically significant. From the regression output above, leaving all other variables fixed, we estimate that the logarithm of relative root median square error in estimating the 95^{th} percentile of lognormal data : (i) increases by 2.710 per unit increase in censoring rate; (ii) is quadratic in sigma, first increasing and then decreasing and (iii) is less than the correct lognormal analysis when the incorrect Weibull analysis is used.

CHAPTER 6 - Conclusions

The larger the sample size and the lower the censoring rate the better the performance of the estimates of the 5^{th} percentile of Weibull data. But, we observed some cases where the performances of the estimates got worse with an increase in sample size when estimating the 95^{th} percentile.

For the smaller sample sizes in my study, both correct and incorrect parametric estimates of the 5^{th} percentile perform better than the nonparametric Kaplan-Meier estimates for Weibull data. For the larger sample sizes, the nonparametric Kaplan-Meier estimates of the 5^{th} percentile were closer to the correct parametric estimates. Further, the nonparametric Kaplan-Meier estimates of the 95^{th} percentile were in many cases even better than the correct parametric estimates.

The larger the shape parameter of the Weibull data, the better the performance of both 5^{th} and 95^{th} percentile estimates.

Although the Parametric estimates of 5^{th} percentile are robust with respect to the assumed underlying distribution, they tend to become less robust as we move from estimating the 5^{th} toward the 95^{th} percentile for Weibull data.

The nonparametric Kaplan-Meier estimates performed best as the censoring rate increases with Weibull data. In contrast, the parametric Weibull estimates did the best with lognormal data.

For lower censoring rates the performance of incorrect parametric estimates of the 5^{th} percentile remain stable with lognormal data. It demonstrated the same behavior as with Weibull data only for higher censoring rates.

The lognormal data are very sensitive to the censoring rate and we observed that for higher censoring rates the incorrect parametric estimates perform the best with lognormal data .

If you do not know the underlying distribution of the data, it is risky to use the parametric estimates of quantiles close to one. A limitation in using the nonparametric estimates is the possibly high proportion of data sets for which an estimate of a large quantile is not available when the censoring rate is large. For future work we suggest that in such cases, always designate the largest observation as being uncensored and then find the nonparametric Kaplan-Meier estimates.

References

John P. Klein, Melvin L. Moeschberger (2003). Survival Analysis. Techniques for Censored and Truncated Data, Second Edition. Springer-Verlag, New York. (pp. 19-29).

Cantor, Alan B. (2003). SAS Survival Analysis Techniques for Medical Research, Second Edition, Cary, NC: SAS Institute Inc. (pp. 18-20).

Engineering Statistics Handbook at
[http://www.itl.nist.gov/div898/handbook/index.htm]

[http://www.engineeredsoftware.com/nasa/weibull.htm]

[http://www.engineeredsoftware.com/nasa/Lognormal.htm]

[http://en.wikipedia.org/wiki/Box%E2%80%93Muller_transform]

[http://www.weibull.com/LifeDataWeb/characteristics_of_the_weibull_distribution.htm]

[http://support.sas.com/documentation/cdl/en/lrdict/63026/HTML/default/viewer.htm#/documentation/cdl/en/lrdict/63026/HTML/default/a000202883.htm]

Cody Ron, (2010), SAS Functions by example, Second Edition, SAS Press

Cody Ron, (2007), Learning SAS by examples: A Programmer's Guide, SAS Press

Appendix A - SAS Code and Output for a Real Data Set

SAS Code - Largest observation censored

```
data aneuploid;
input DT Censor @@;
datalines;
1     1     3     1     3     1     4     1     10    1
13    1     13    1     16    1     16    1     24    1
26    1     27    1     28    1     30    1     30    1
32    1     41    1     51    1     65    1     67    1
70    1     72    1     73    1     77    1     91    1
93    1     96    1     100   1     104   1     157   1
167   1     61    0     74    0     79    0     80    0
81    0     87    0     87    0     88    0     89    0
93    0     97    0     101   0     104   0     108   0
109   0     120   0     131   0     150   0     231   0
240   0     400   0
;
run;
proc print;
run;

proc lifetest data=aneuploid method=KM;
TIME DT*Censor(0);
run;

proc lifereg data=aneuploid;
model DT*Censor(0)= / dist=Weibull;
run;

proc lifereg data=aneuploid;
model DT*Censor(0)= / dist=LNormal;
run;
```

Output of Proc Lifetest

Product-Limit Survival Estimates

DT	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	52
1.000	0.9808	0.0192	0.0190	1	51
3.000	.	.	.	2	50
3.000	0.9423	0.0577	0.0323	3	49
4.000	0.9231	0.0769	0.0370	4	48
10.000	0.9038	0.0962	0.0409	5	47
13.000	.	.	.	6	46
13.000	0.8654	0.1346	0.0473	7	45
16.000	.	.	.	8	44
16.000	0.8269	0.1731	0.0525	9	43
24.000	0.8077	0.1923	0.0547	10	42
26.000	0.7885	0.2115	0.0566	11	41
27.000	0.7692	0.2308	0.0584	12	40
28.000	0.7500	0.2500	0.0600	13	39
30.000	.	.	.	14	38
30.000	0.7115	0.2885	0.0628	15	37
32.000	0.6923	0.3077	0.0640	16	36
41.000	0.6731	0.3269	0.0651	17	35
51.000	0.6538	0.3462	0.0660	18	34
61.000*	.	.	.	18	33
65.000	0.6340	0.3660	0.0669	19	32
67.000	0.6142	0.3858	0.0677	20	31
70.000	0.5944	0.4056	0.0683	21	30
72.000	0.5746	0.4254	0.0689	22	29
73.000	0.5548	0.4452	0.0693	23	28
74.000*	.	.	.	23	27
77.000	0.5342	0.4658	0.0697	24	26
79.000*	.	.	.	24	25
80.000*	.	.	.	24	24
81.000*	.	.	.	24	23
87.000*	.	.	.	24	22
87.000*	.	.	.	24	21
88.000*	.	.	.	24	20
89.000*	.	.	.	24	19
91.000	0.5061	0.4939	0.0715	25	18
93.000	0.4780	0.5220	0.0728	26	17
93.000*	.	.	.	26	16
96.000	0.4481	0.5519	0.0741	27	15
97.000*	.	.	.	27	14
100.000	0.4161	0.5839	0.0754	28	13
101.000*	.	.	.	28	12
104.000	0.3814	0.6186	0.0767	29	11
104.000*	.	.	.	29	10
108.000*	.	.	.	29	9
109.000*	.	.	.	29	8

120.000*	.	.	.	29	7
131.000*	.	.	.	29	6
150.000*	.	.	.	29	5

The LIFETEST Procedure

Product-Limit Survival Estimates

DT	Survival	Failure	Survival Standard Error	Number Failed	Number Left
157.000	0.3051	0.6949	0.0918	30	4
167.000	0.2289	0.7711	0.0954	31	3
231.000*	.	.	.	31	2
240.000*	.	.	.	31	1
400.000*	.	.	.	31	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable DT

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	167.000	104.000	.
50	93.000	67.000	157.000
25	29.000	16.000	67.000

Mean	Standard Error
93.320	9.222

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
52	31	21	40.38

Output of Proc Lifereg with dist=Weibull

The LIFEREG Procedure

Model Information

Data Set	WORK.ANEUPLOID
Dependent Variable	Log(DT)
Censoring Variable	Censor
Censoring Value(s)	0
Number of Observations	52
Noncensored Values	31
Right Censored Values	21
Left Censored Values	0
Interval Censored Values	0
Name of Distribution	Weibull
Log Likelihood	-76.35881313

Number of Observations Read	52
Number of Observations Used	52

Algorithm converged.

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.9604	0.2219	4.5254	5.3953	499.65	<.0001
Scale	1	1.2017	0.1847	0.8891	1.6242		
Weibull Scale	1	142.6472	31.6553	92.3359	220.3717		
Weibull Shape	1	0.8322	0.1279	0.6157	1.1248		

Output of Proc Lifereg with dist=Lognormal

The LIFEREG Procedure

Model Information

Data Set	WORK.ANEUPLOID
Dependent Variable	Log(DT)
Censoring Variable	Censor
Censoring Value(s)	0
Number of Observations	52
Noncensored Values	31
Right Censored Values	21
Left Censored Values	0
Interval Censored Values	0
Name of Distribution	Lognormal
Log Likelihood	-76.42406026

Number of Observations Read	52
Number of Observations Used	52

Algorithm converged.

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.4633	0.2702	3.9338	4.9927	272.95	<.0001
Scale	1	1.7150	0.2342	1.3122	2.2414		

SAS Code-Largest observation uncensored

```
data aneuploid;
input DT Censor @@;
datalines;
1     1     3     1     3     1     4     1     10    1
13    1     13    1     16    1     16    1     24    1
26    1     27    1     28    1     30    1     30    1
32    1     41    1     51    1     65    1     67    1
70    1     72    1     73    1     77    1     91    1
93    1     96    1     100   1     104   1     157   1
167   1     61    0     74    0     79    0     80    0
81    0     87    0     87    0     88    0     89    0
93    0     97    0     101   0     104   0     108   0
109   0     120   0     131   0     150   0     231   0
240   0     400   1
;
run;
proc print;
run;

proc print;
run;

title 'Plot of Kaplan-Meier Estimator';
title2 'when the largest observation is uncensored';

symbol1 interpol=join width=1 color=red value=dot height=1;

proc lifetest data=aneuploid method=KM plots = (s,lls);
TIME DT*Censor(0);
run;

proc lifereg data=aneuploid;
model DT*Censor(0)= / dist=Weibull;
run;

proc lifereg data=aneuploid;
model DT*Censor(0)= / dist=LNormal;
run;
```

Output of Proc Lifetest

The SAS System 14:34 Wednesday, March 31, 2010 1

Obs	DT	Censor
1	1	1
2	3	1
3	3	1
4	4	1
5	10	1
6	13	1
7	13	1
8	16	1
9	16	1
10	24	1
11	26	1
12	27	1
13	28	1
14	30	1
15	30	1
16	32	1
17	41	1
18	51	1
19	65	1
20	67	1
21	70	1
22	72	1
23	73	1
24	77	1
25	91	1
26	93	1
27	96	1
28	100	1
29	104	1
30	157	1
31	167	1
32	61	0
33	74	0
34	79	0
35	80	0
36	81	0
37	87	0
38	87	0
39	88	0
40	89	0
41	93	0
42	97	0
43	101	0
44	104	0
45	108	0
46	109	0
47	120	0
48	131	0
49	150	0

50	231	0
51	240	0
52	400	1

The LIFETEST Procedure

Product-Limit Survival Estimates

DT	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	52
1.000	0.9808	0.0192	0.0190	1	51
3.000	.	.	.	2	50
3.000	0.9423	0.0577	0.0323	3	49
4.000	0.9231	0.0769	0.0370	4	48
10.000	0.9038	0.0962	0.0409	5	47
13.000	.	.	.	6	46
13.000	0.8654	0.1346	0.0473	7	45
16.000	.	.	.	8	44
16.000	0.8269	0.1731	0.0525	9	43
24.000	0.8077	0.1923	0.0547	10	42
26.000	0.7885	0.2115	0.0566	11	41
27.000	0.7692	0.2308	0.0584	12	40
28.000	0.7500	0.2500	0.0600	13	39
30.000	.	.	.	14	38
30.000	0.7115	0.2885	0.0628	15	37
32.000	0.6923	0.3077	0.0640	16	36
41.000	0.6731	0.3269	0.0651	17	35
51.000	0.6538	0.3462	0.0660	18	34
61.000*	.	.	.	18	33
65.000	0.6340	0.3660	0.0669	19	32
67.000	0.6142	0.3858	0.0677	20	31
70.000	0.5944	0.4056	0.0683	21	30
72.000	0.5746	0.4254	0.0689	22	29
73.000	0.5548	0.4452	0.0693	23	28
74.000*	.	.	.	23	27
77.000	0.5342	0.4658	0.0697	24	26
79.000*	.	.	.	24	25
80.000*	.	.	.	24	24
81.000*	.	.	.	24	23
87.000*	.	.	.	24	22
87.000*	.	.	.	24	21
88.000*	.	.	.	24	20
89.000*	.	.	.	24	19
91.000	0.5061	0.4939	0.0715	25	18
93.000	0.4780	0.5220	0.0728	26	17
93.000*	.	.	.	26	16
96.000	0.4481	0.5519	0.0741	27	15
97.000*	.	.	.	27	14
100.000	0.4161	0.5839	0.0754	28	13
101.000*	.	.	.	28	12
104.000	0.3814	0.6186	0.0767	29	11

104.000*	.	.	.	29	10
108.000*	.	.	.	29	9
109.000*	.	.	.	29	8
120.000*	.	.	.	29	7
131.000*	.	.	.	29	6
150.000*	.	.	.	29	5

The LIFETEST Procedure

Product-Limit Survival Estimates

DT	Survival	Failure	Survival Standard Error	Number Failed	Number Left
157.000	0.3051	0.6949	0.0918	30	4
167.000	0.2289	0.7711	0.0954	31	3
231.000*	.	.	.	31	2
240.000*	.	.	.	31	1
400.000	0	1.0000	0	32	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable DT

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	167.000	104.000	400.000
50	93.000	67.000	157.000
25	29.000	16.000	67.000

Mean	Standard Error
146.645	28.105

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
52	32	20	38.46

Output of Proc Lifereg with dist=Weibull

The LIFEREG Procedure

Model Information

Data Set	WORK.ANEUPLOID
Dependent Variable	Log(DT)
Censoring Variable	Censor
Censoring Value(s)	0
Number of Observations	52
Noncensored Values	32
Right Censored Values	20
Left Censored Values	0
Interval Censored Values	0
Name of Distribution	Weibull
Log Likelihood	-75.61481254

Number of Observations Read	52
Number of Observations Used	52

Algorithm converged.

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.9083	0.2059	4.5048	5.3119	568.27	<.0001
Scale	1	1.1435	0.1712	0.8527	1.5334		
Weibull Scale	1	135.4148	27.8820	90.4489	202.7351		
Weibull Shape	1	0.8745	0.1309	0.6521	1.1728		

Output of Proc Lifereg with dist=Lognormal

The SAS System 08:57 Tuesday, June 15, 2010 7

The LIFEREG Procedure

Model Information

Data Set	WORK.ANEUPLOID
Dependent Variable	Log(DT)
Censoring Variable	Censor
Censoring Value(s)	0
Number of Observations	52
Noncensored Values	32
Right Censored Values	20
Left Censored Values	0
Interval Censored Values	0
Name of Distribution	Lognormal
Log Likelihood	-76.57023664

Number of Observations Read	52
Number of Observations Used	52

Fit Statistics

-2 Log Likelihood	153.140
AIC (smaller is better)	157.140
AICC (smaller is better)	157.385
BIC (smaller is better)	161.043

Algorithm converged.

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.4242	0.2592	3.9163	4.9322	291.44	<.0001
Scale	1	1.6630	0.2214	1.2811	2.1588		

Appendix B - Relative Root Mean Square Error and Relative Bias Plots

Plots for Weibull Data

Figure B.1 Relative Root Mean Square Error Plots for Weibull Data with $\hat{a} = 0.5$ and $\ddot{e} = 1$

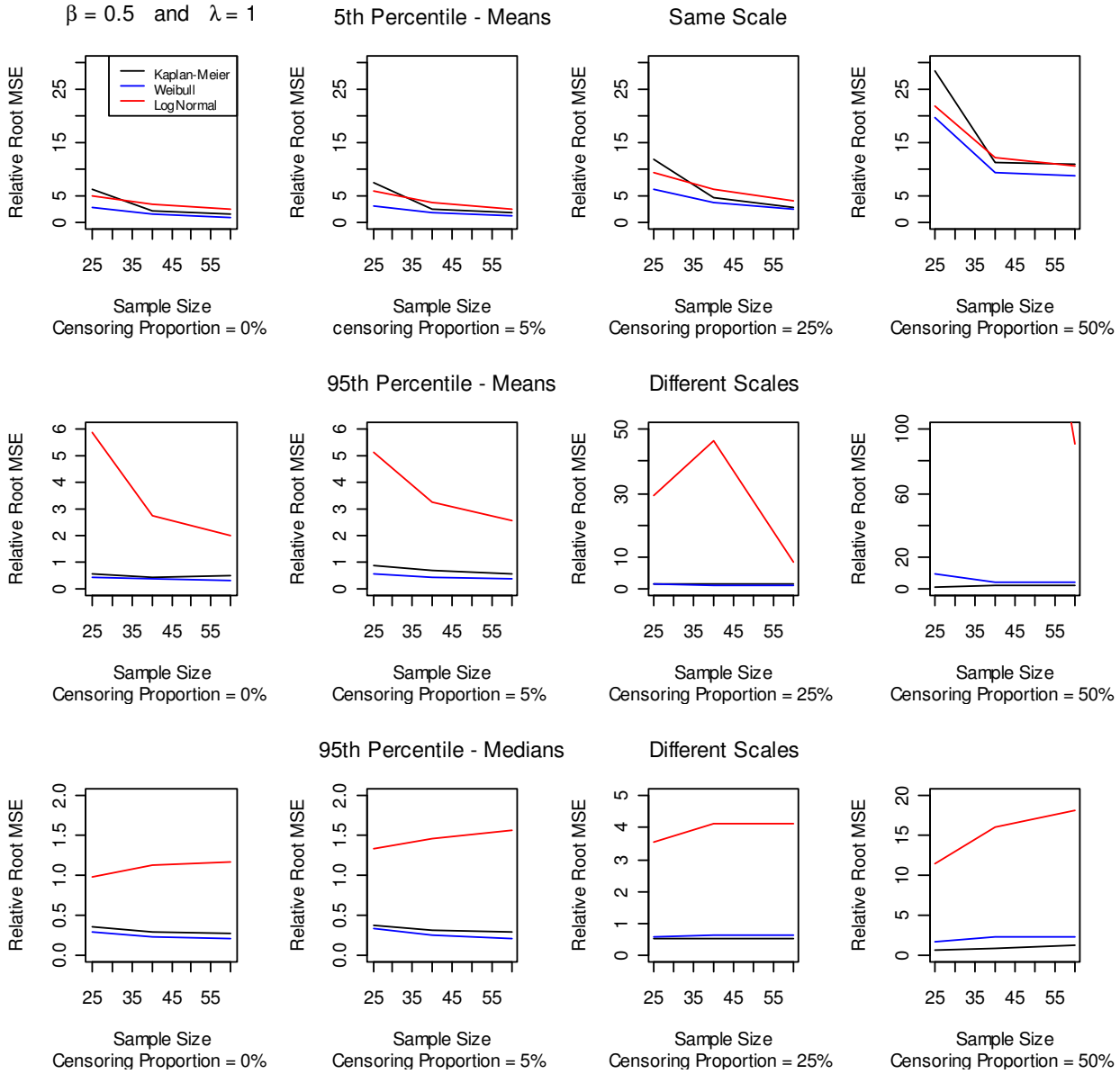


Figure B.2 Relative Bias Plots for Weibull Data with $\hat{a} = 0.5$ and $\tilde{e} = 1$

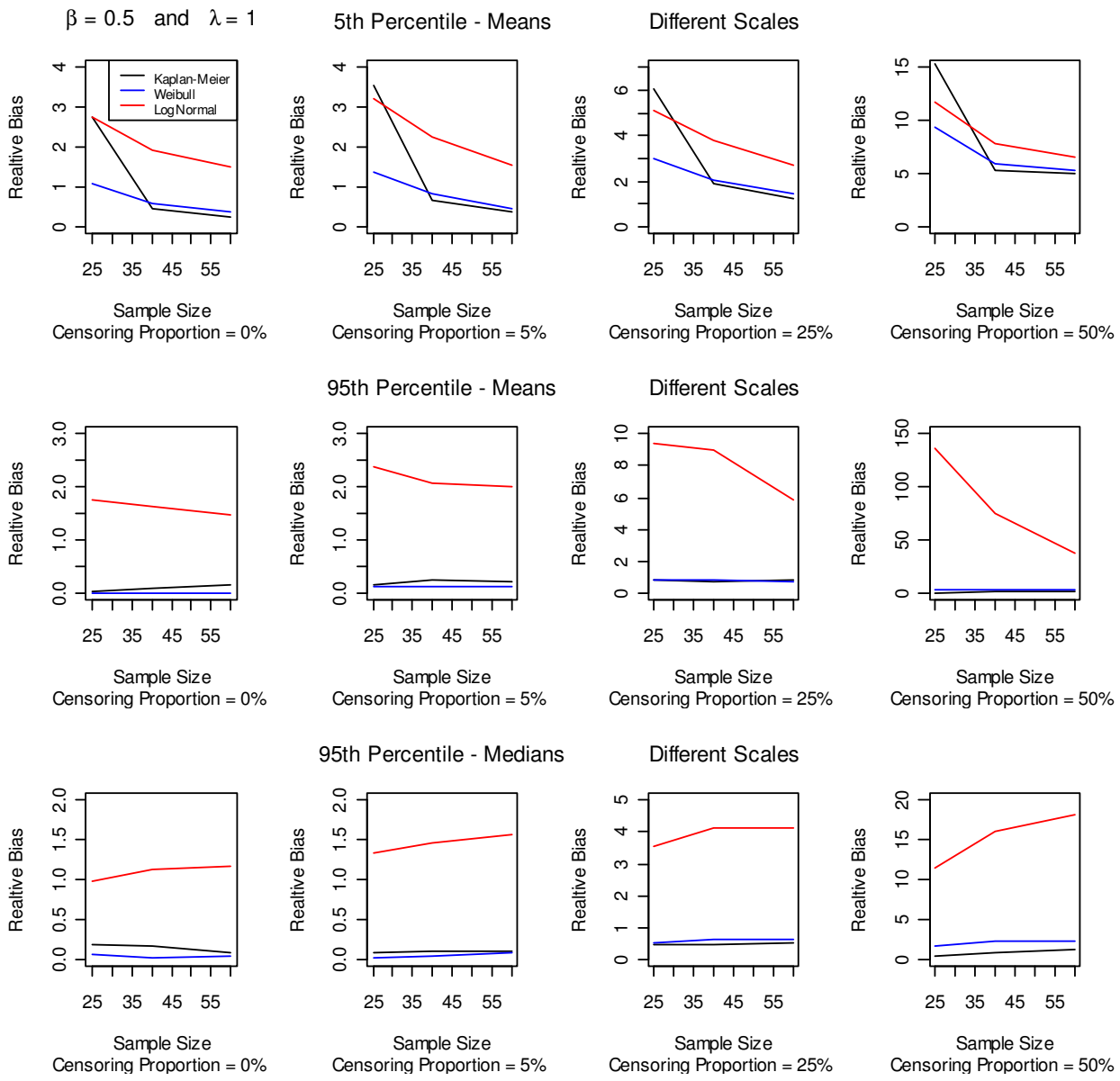


Figure B.3 Relative Root Mean Square Error Plots for Weibull Data with $\hat{a} = 1.5$ and $\ddot{e} = 1$

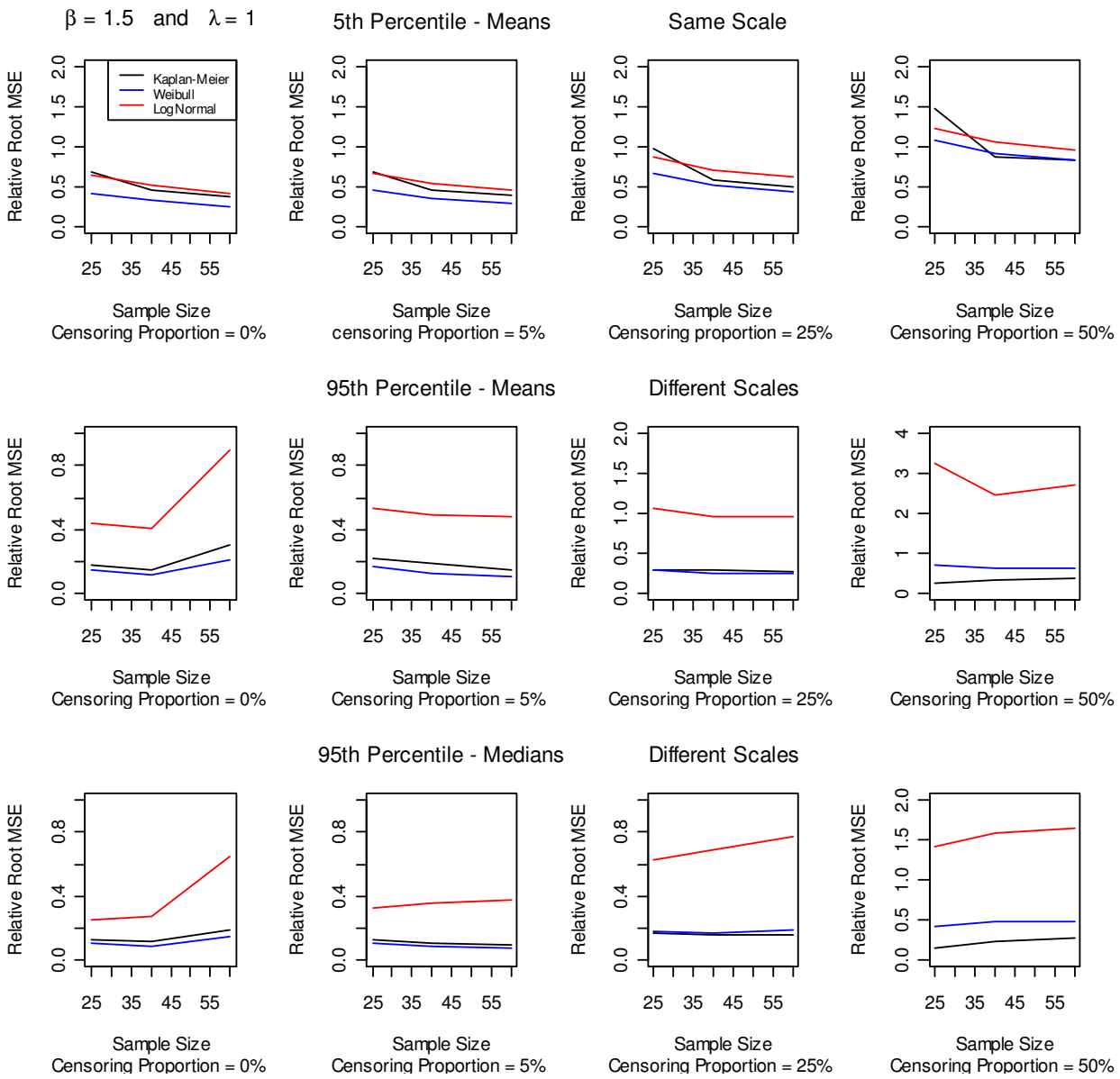


Figure B.4 Relative Bias Plots for Weibull Data with $\hat{a} = 1.5$ and $\tilde{e} = 1$

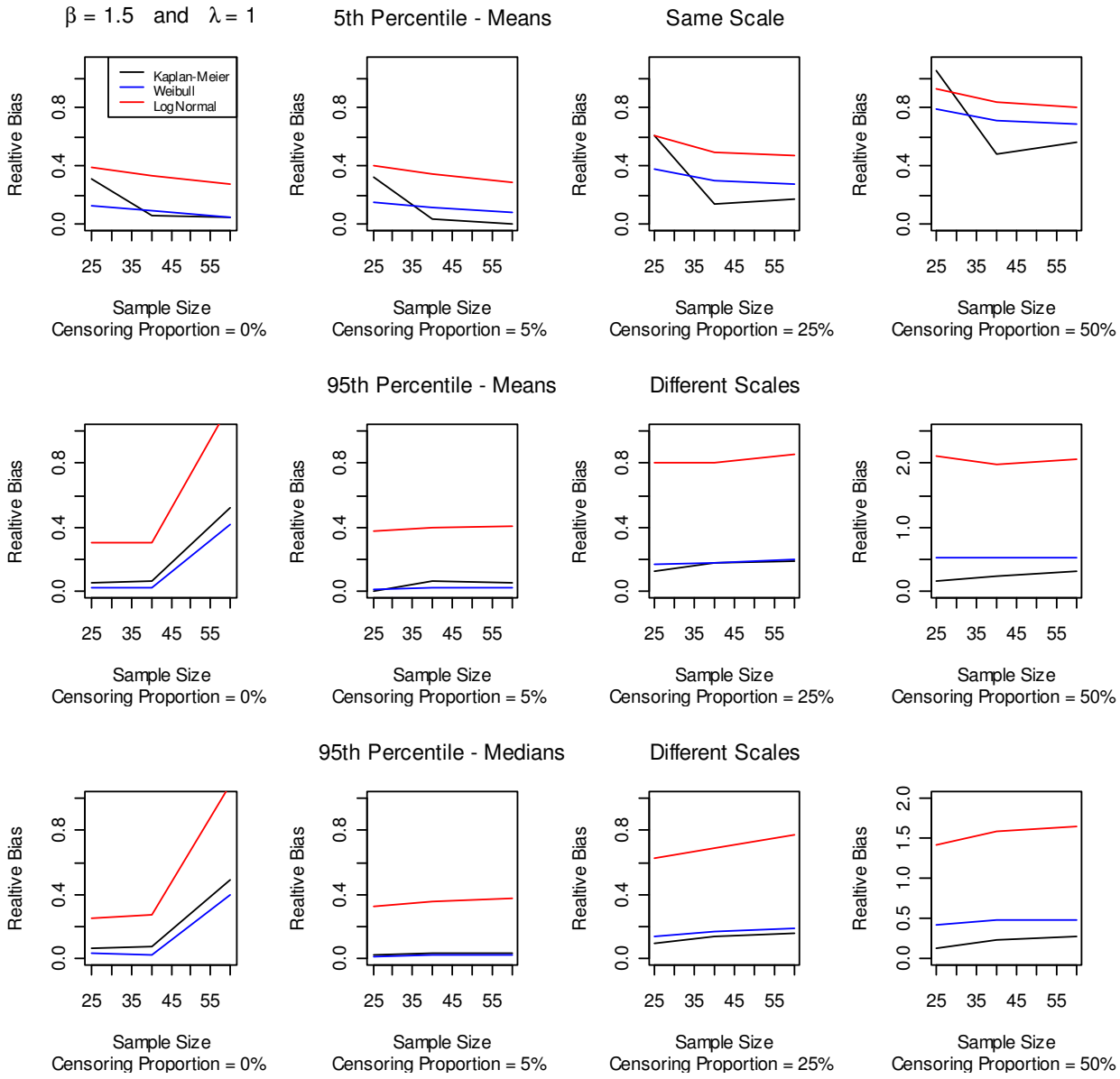


Figure B.5 Relative Root Mean Square Error Plots for Weibull Data with $\hat{a} = 5.0$ and $\ddot{e} = 1$

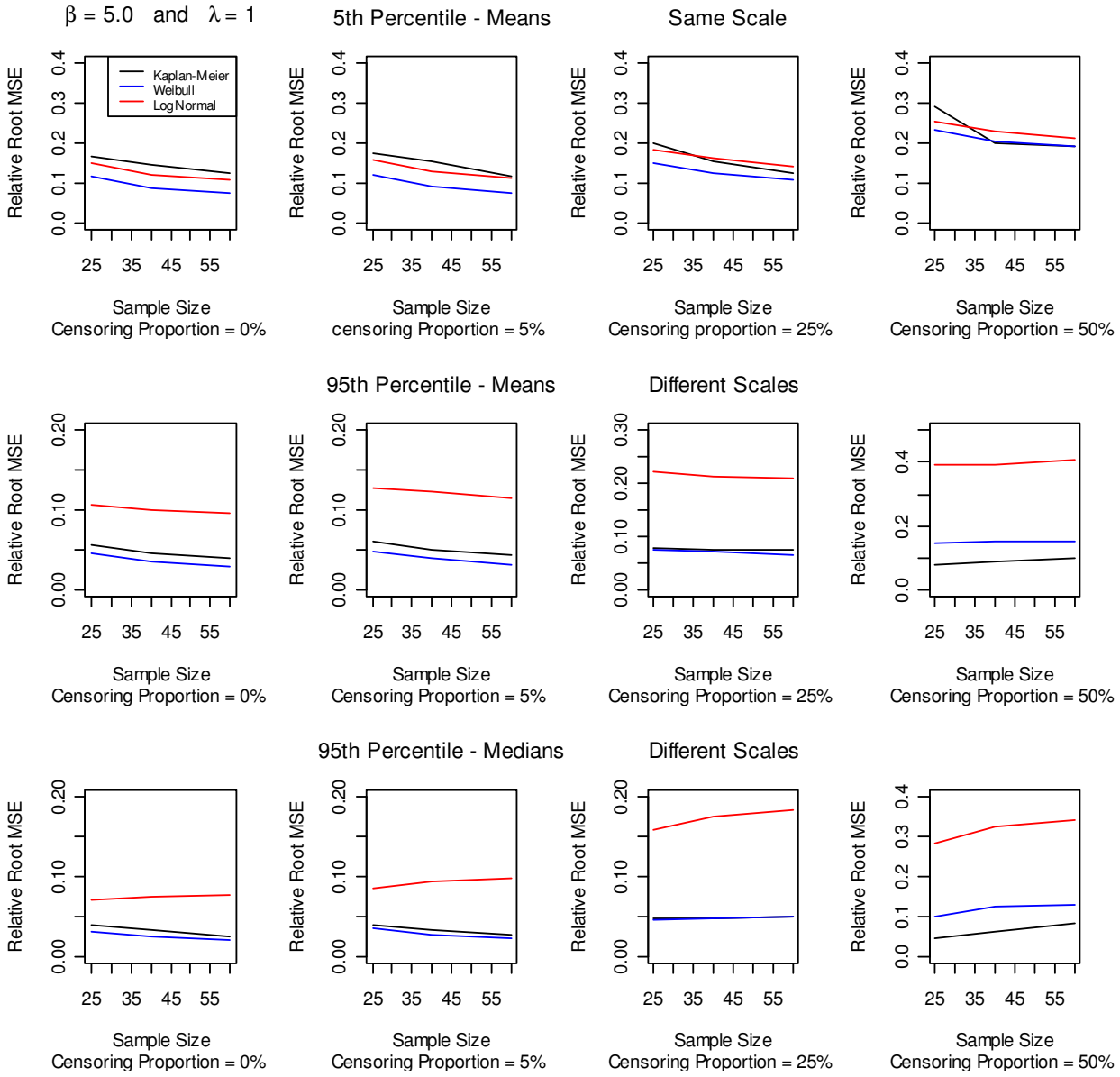
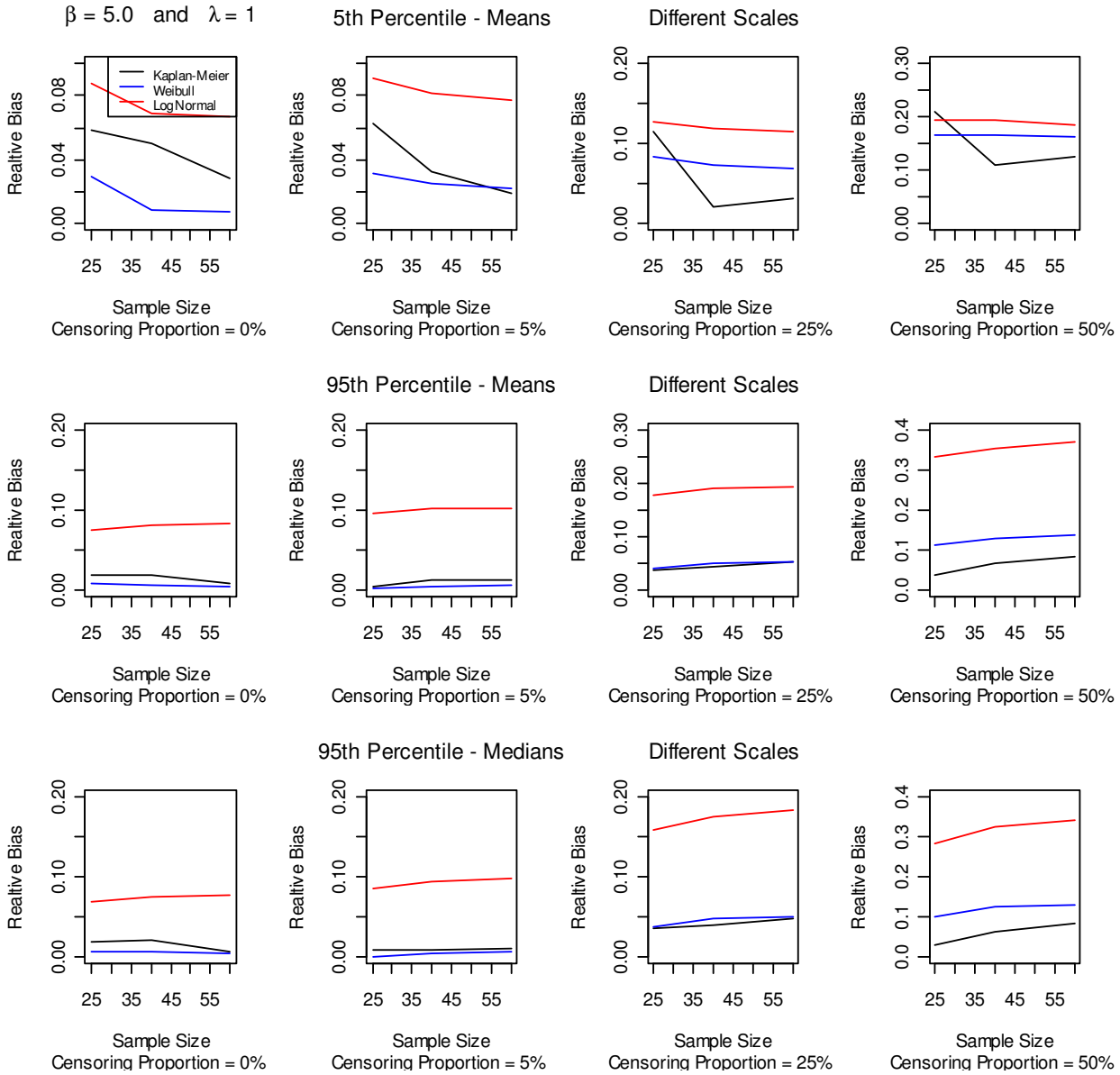


Figure B.6 Relative Bias Plots for Weibull Data with $\beta = 5.0$ and $\lambda = 1$



Plots for Lognormal Data

Figure B.7 Relative Root Mean Square Error Plots for Lognormal Data with ó = 1.0 and ì = 0

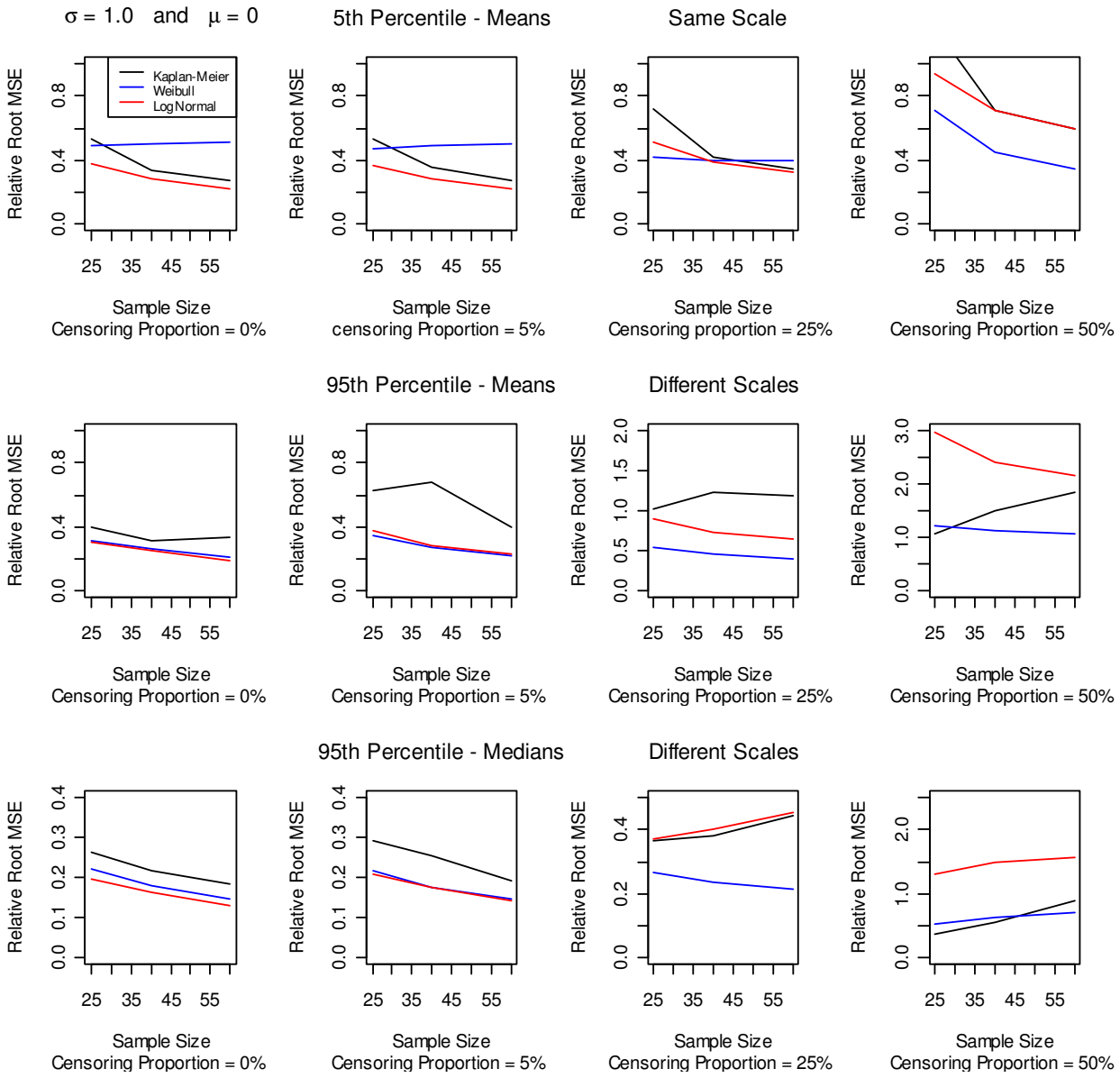


Figure B.8 Relative Bias Plots for Lognormal Data with ó = 1. 0 and ì = 0

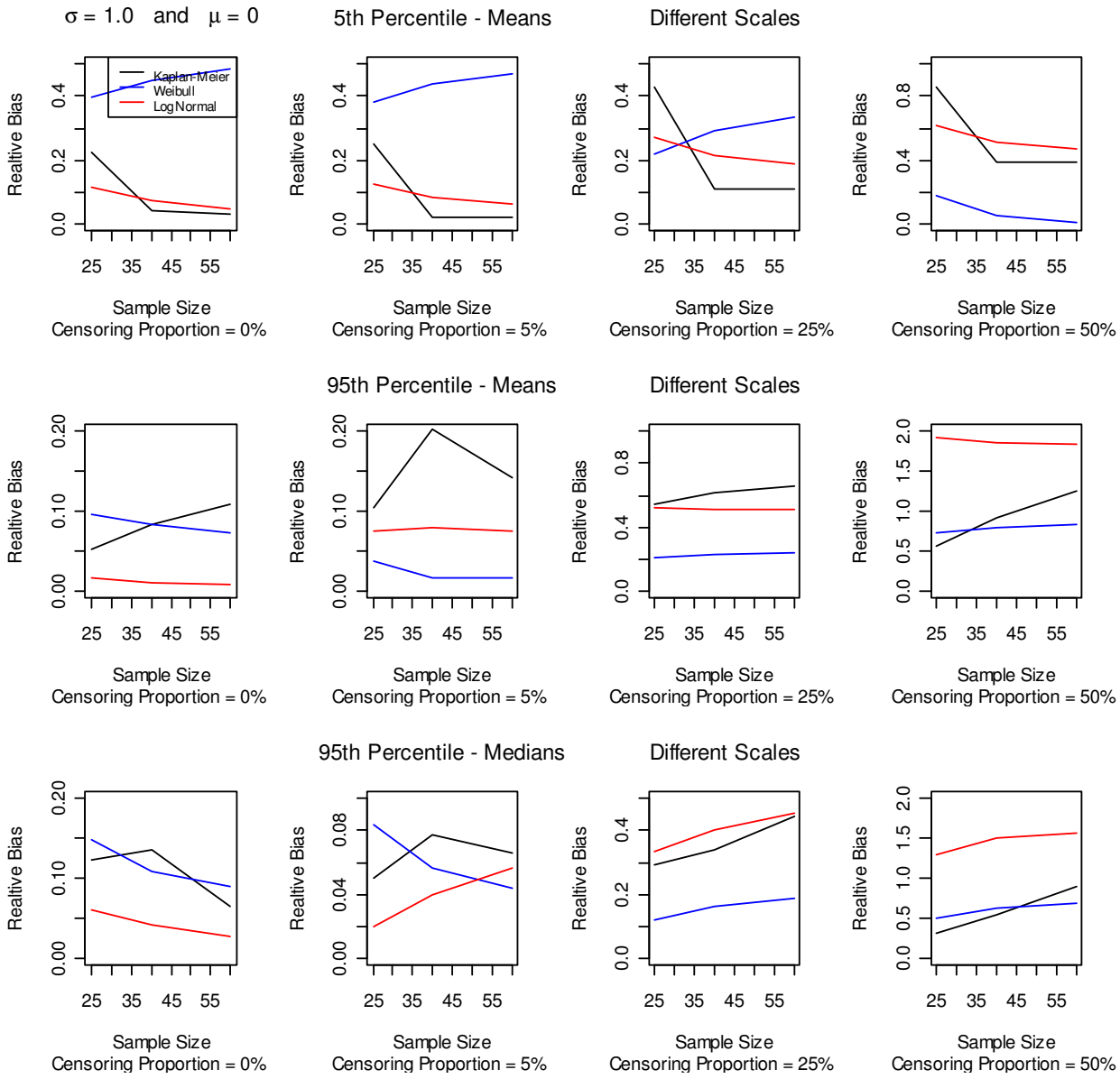


Figure B.9 Relative Root Mean Square Error Plots for Lognormal Data with $\sigma = 1.5$ and $\mu = 0$

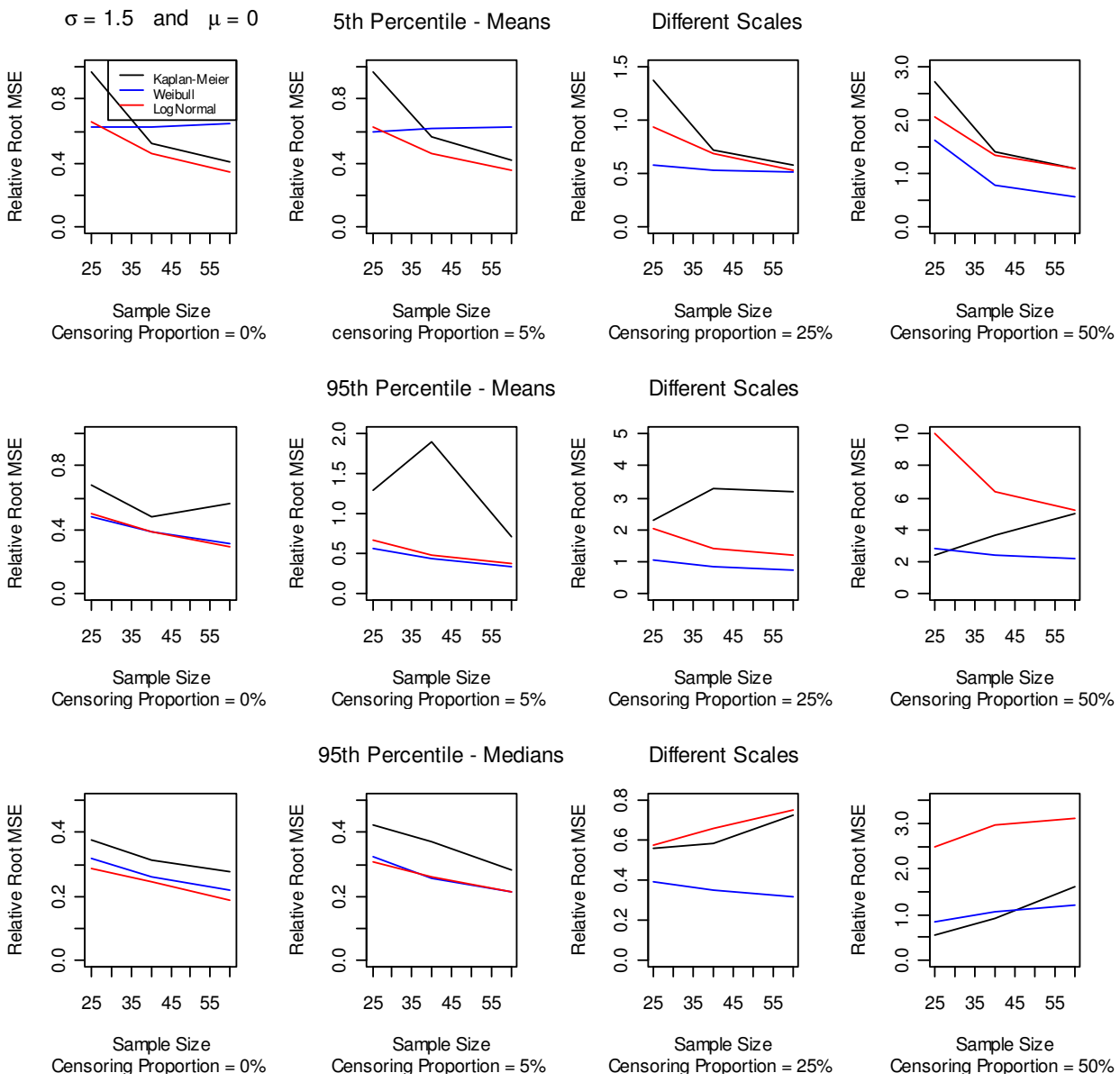


Figure B.10 Relative Bias Plots for Lognormal Data with ó = 1. 5 and ì = 0

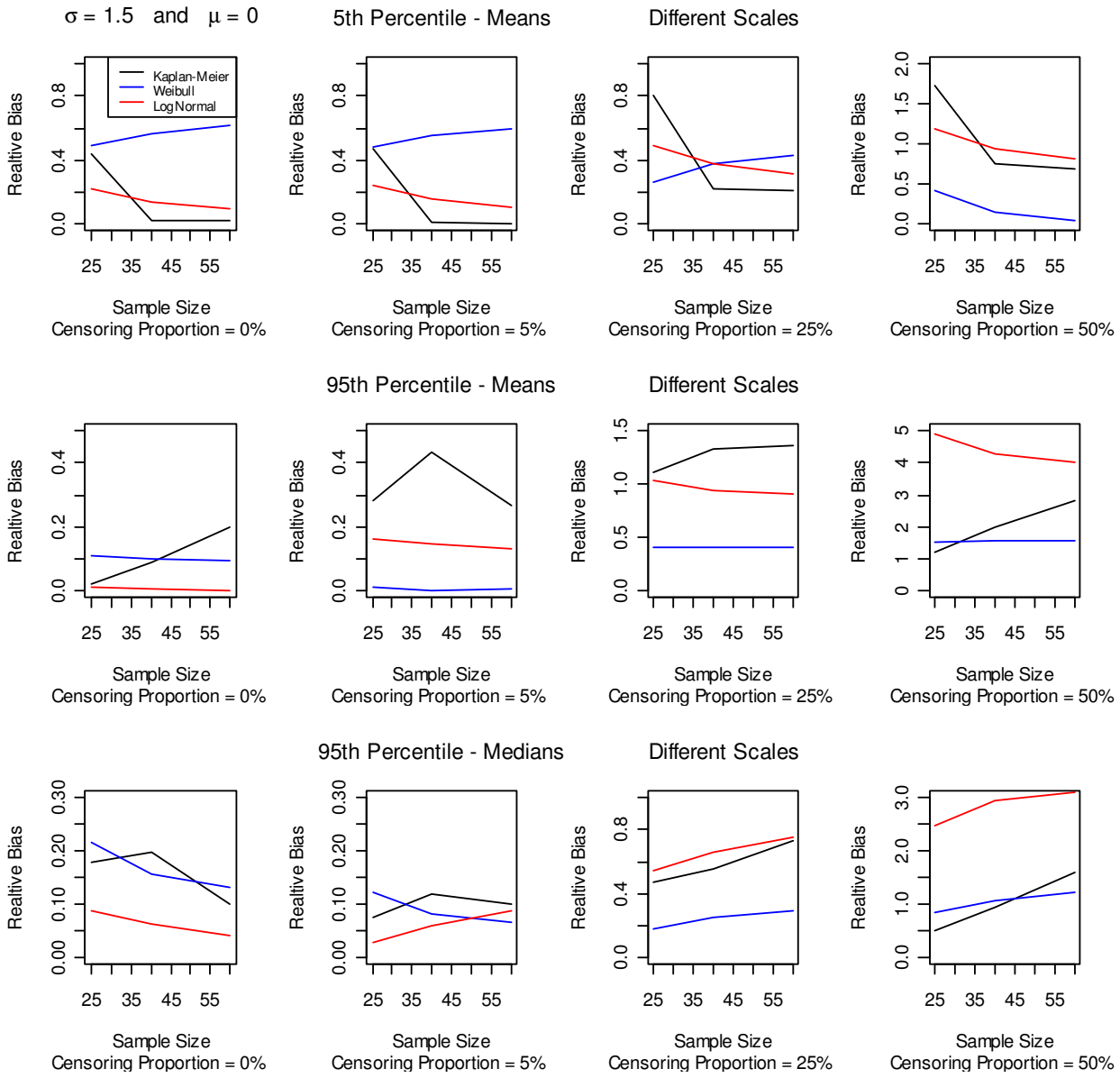


Figure B.11 Relative Root Mean Square Error Plots for Lognormal Data with ó = 10 and ì = 0

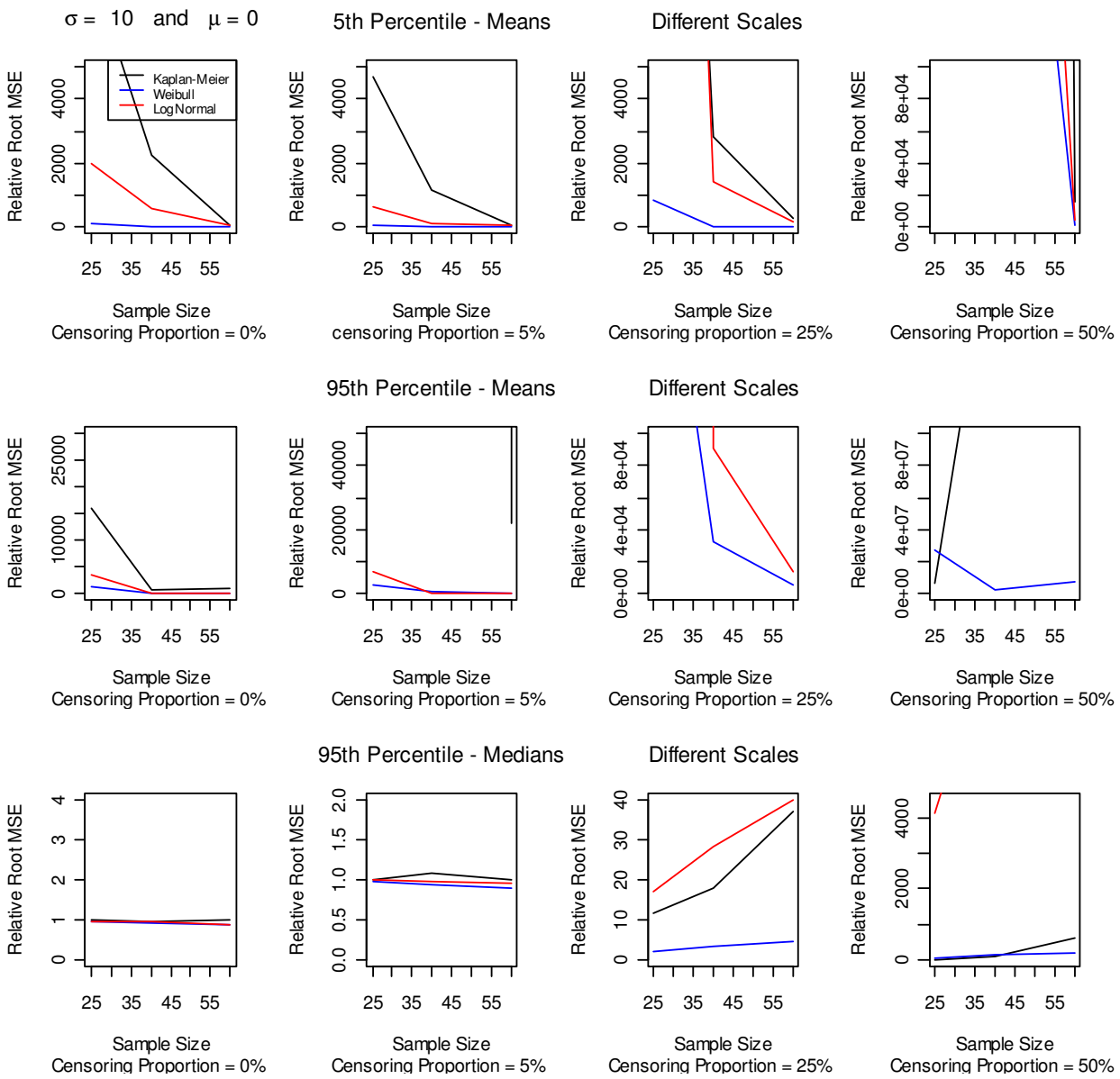


Figure B.12 Relative Bias Plots for Lognormal Data with ó = 10 and ì = 0